Alternative data sources in the insurance industry

Institute of Actuaries of India

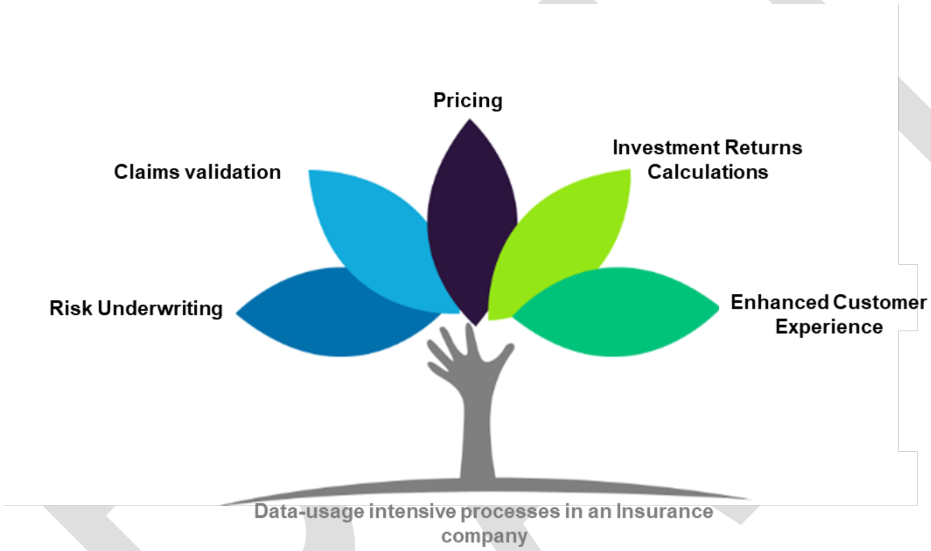
INDIA INSURTECH ASSOCIATION

Exposure Draft of Whitepaper authored by joint working party of Institute of Actuaries of India and India Insurtech Association

Table of contents

1. Executive summary

The insurance sector plays a crucial role in fostering sustainable economic development within a nation. As the number of individuals purchasing insurance policies continues to rise, it has become imperative for insurance players to establish comprehensive systems to manage and analyse data from alternative sources and to use alternative data sources in their regular processes. All major processes like risk underwriting, claims validation, pricing, investment returns calculation and enhanced customer experience require intensive data usage.


Pricing
Claims validation
Investment Returns Calculations
Risk Underwriting
Enhanced Customer Experience
Data-usage intensive processes in an Insurance company

Alternative data source applications in the insurance industry

Undoubtedly, data has emerged as a pivotal factor that sets industry players apart. Those with greater access to reliable data and the ability to effectively process it possess a distinct advantage over their competitors.

Furthermore, the insurance industry is undergoing rapid transformation through digitalisation and the emergence of InsurTech firms. These changes have given rise to novel sources of customer data that insurance companies can leverage to enhance various processes within the value chain, making it more efficient and effective. This data can help insurers with internal analysis and to attempt to solve specific challenges, innovate new products or services, and improve existing processes within the insurance ecosystem.

Although there has been a rise in the use of alternative data sources in various processes of insurance players such as insurers, re-insurers, brokers, intermediaries and InsurTech firms, extraction of data from sources and analysis poses various challenges:

a. incomplete or unverifiable data
b. unstructured data leading to difficulties in integration

c. no permission for redistribution of data
d. limited use of data.

We have developed a framework that helps to mitigate these challenges by rating data sources on aspects such as credibility, ease of automation, recency, pace of processing and cost involved. The framework provides a clear idea to an insurance company about whether a particular data source may be used in their day-to-day processes on a sustainable and long-term basis.

2. Background and purpose

Alternative data is publicly and privately available data that can be gathered responsibly. Traditionally, in the insurance industry, the data is gathered at source systems. Information on the policyholders, credit scores, claims history and sum assured is traditionally collected at the time of policy inception. Alternative data, on the other hand, refers to non-conventional or non-traditional data that insurers can utilise for underwriting policies, assessing risks and making well-informed decisions. Moreover, this data provides additional insights into the policyholder's habits, behaviour and preferences.

Insurance players can utilise alternative data to improve the efficiency of their value chain processes (underwriting, claims, pricing, etc.), leading to:

- accurate and informed decisions
- quicker TATs
- better customer satisfaction
- lower customer acquisition costs (CACs)
- higher market share.

Alternative datasets are changing the landscape for companies operating in the insurance sector. Although it is more relevant to general insurers, alternative data can also prove useful to life insurers, distributors and InsurTech firms for lead acquisition, underwriting, claims management and risk assessment.

This report aims to provide a framework that can help the insurance industry enhance the insurance value chain by:

- identifying and evaluating alternative data sources available in India based on their use cases, including qualitative and quantitative parameters
- leveraging alternative data sources as part of their processes in the entire value-chain for e.g. claims settlement and underwriting.
- publishing a regularly updated repository of alternative data sources by crowdsourcing inputs from the insurance industry eco-system.

3. Some use cases of alternative data in the insurance industry

3.1 **Lemonade** is a new age insurance company established in 2016. Lemonade collects more than 100 data points per customer to speed up claim settlements, personalise coverage and improve overall customer satisfaction. To do so, Lemonade uses the following tools:

a. AI chatbots: Lemonade uses AI-based chatbots to gather information, onboard clients, underwrite and settle claims. Maya is a bot designed to provide superior customer experience to the end user, analysing vast amounts of external and internal data to make situational decisions in real time. The company uses a bot named Jim for claims settlement.[1]

b. IoT sensor data: By integrating with IoT devices like smart home devices and security systems, Lemonade accesses real-time data on home security, occupancy patterns and potential risks.[2]

How does alternative data help Lemonade?

a. Occupancy patterns: IoT sensors can provide insights into occupancy patterns within a property. By analysing data on motion sensors, door or window sensors and connected devices, Lemonade can determine when a property is typically occupied or vacant. This information helps in understanding risk factors related to burglaries, fire hazards and property damage.

b. **Personalised pricing and risk assessment:** By accessing real-time data from IoT devices, Lemonade can offer more personalised pricing and risk assessment. Policyholders who have implemented advanced security systems or smart home devices may be eligible for lower premiums based on reduced risks.

c. Proactive risk mitigation: Data from IoT devices (e.g. sensors) allows the company to detect risks at an advanced stage and devise a strategy to avoid them. For instance, if there's a fire at an insured property, the information from the IoT device will help manage the risk from an early stage, leading to lesser damage to the property, thus resulting in a lower claim amount.

d. Enhanced claims process: In the event of an unfortunate event like fire or flooding, IoT devices would share accurate information with the insurers in real time, leading to accurate claim estimation and quicker settlement.

[1] https://www.forbes.com/sites/garydrenik/2022/09/27/how-ai-is-changing-the-game-in-insurance/?sh=ce905cf51bf5

[2] https://edition.cnn.com/2021/05/27/tech/lemonade-ai-insurance/index.html

3.2 **Hippo** is a new-age insurance company primarily focussed on property insurance. It utilises multiple data points for analysis to get a comprehensive picture about the properties it has insured. Below are a few alternative data sources that it leverages in its processes:

   a. **Public records:** Hippo gathers publicly available information – permits of a building, tax-related information, etc. – to get a comprehensive view of the insurance property. These records are helpful in getting insights like the date when the construction was done, date of the last renovation, total area covered etc.

   b. **Geospatial data:** By accessing geospatial data, the company is able to determine if the property is located in risk-prone areas – flood zone, nearby areas where other hazards like cyclone or tsunami are common, etc. – which affects the risk scoring of the property.

   c. **Weather data**: By integrating weather-related data into their processes, the company is able to verify if the property is more prone to weather-related risks like cyclones, storms, etc.

   d. **Smart home devices**: By integrating IoT devices with the insured properties, Hippo is able to detect any unfortunate event early on and any potential risks associated with the property.[3]

How does alternative data help Hippo Insurance?

a. **Enhanced risk assessment:** Using data from alternative data sources, Hippo gains much deeper insights into the insured property, leading to more comprehensive underwriting that factor in the risks associated with the property's construction details, location and other related environmental factors. This helps in offering customised pricing to end users.

b. **Customised coverage:** By using data from alternate data sources, Hippo can offer customised coverage to all its clients. For instance, if the property to be insured is equipped with state-of-the-art alert and security systems or IoT devices, then the risks of robbery or damage by natural factors is reduced, leading to the company offering more favourable terms to end users.

c. **Proactive risk mitigation:** By using data from alternative data sources like weather data, Hippo can alert the property owners/tenants to be better prepared for any unusual event. This leads to lesser damage and hence significant reduction in claim payouts.[4]

d. **Streamlined claims process:** Data from alternative sources like IoT devices will inform Hippo, in real time, of any damages done to the property which will help in quicker claims validation and settlement.

e. **Improved underwriting accuracy:** By leveraging data from alternative data sources like public records, weather data and IoT sensors, Hippo is able to perform underwriting in a more efficient manner. This ensures that the right pricing is done, in proportion to the associated underlying risks associated.

f. **Competitive advantage:** Alternative data sources allow Hippo to facilitate quick claims settlement, better underwriting and pricing, and proactive risk mitigation. These

[3] https://www.prnewswire.com/news-releases/hippo-insurance-services-provides-new-customers-with-notion-smart-home-sensors-to-prevent-disasters-and-save-on-insurance-policies-300715723.html

[4] https://www.prnewswire.com/news-releases/hippo-insurance-services-provides-new-customers-with-notion-smart-home-sensors-to-prevent-disasters-and-save-on-insurance-policies-300715723.html

advantages provide Hippo an edge over its competitors who still rely on traditional sources of information for their day-to-day processes.

3.3 Usage-based insurance (UBI) can be explained as insurance (motor in this case) based on usage (driving in this case). A device is installed in the vehicle, providing the real-time status of the vehicle like location, mileage, etc. Along with the telematics device, useful data can be fetched from other sources such as GPS devices and mobile apps. Artificial intelligence and machine learning algorithms can then analyse this data and develop a 'risk score' for the customer, which enables the company to perform effective underwriting and quote appropriate pricing for the relevant insurance policies. Below are some key alternative data sources they can leverage:

a. **Telematics data:** Data from such devices installed in vehicles can help analyse the driving behaviour of the driver by providing information like average speed, braking and acceleration pattern, mileage, and location. This can lead to customised pricing as per the overall driving behaviour of an individual.
b. **GPS data:** This data is used to track the movement and location of the vehicle, which helps the insurer in understanding the driving pattern of the insured person – considering factors like distance travelled, locations covered, frequency of driving, etc.
c. **Mobile apps:** Many insurance companies provide their mobile applications to insures persons, which track the speed, location and acceleration of the vehicle, facilitating companies to provide a customised risk rating and efficient underwriting.
d. **OBD-II port devices:** Insurers are adding smart devices that integrate with OBD-II port of vehicles to obtain various data points like engine performance, driving behaviour, mileage etc.
e. **Mobile phone data:** Some insurers, after taking due permissions from policyholders, can get access to their phone data like mails and texts to gather additional insights required for risk underwriting purpose.
f. **In-car cameras:** In-car cameras can capture the video footage of the vehicle, leading to more insights into the driving behaviour and to understand the root cause of any accident better to help determine if the claim is eligible to be serviced or not.

How does UBI help the insurers?

a. **Accurate risk assessment:** Data points like average speed, mileage and acceleration pattern from various sources such as cabin cameras, mobile phone apps and GPS data can help in performing a comprehensive and accurate risk assessment of the individual, thus leading to better and more accurate pricing.

b. **Incentivising safer driving:** By understanding the driving patterns of policyholders, insurers can provide incentives to policyholders with better driving scores in the form of reduced premiums on renewal, enhanced sum assured etc., leading to a culture of responsible driving.

c. **Efficient claims processing:** Seamless passage of data between the insured vehicle and insurer, leads to efficient and quicker processing of claims reporting and settlement.
d. **Enhanced customer engagement:** UBI will lead to better customer engagement by sharing valuable insights like driving behaviour, car health etc., with policyholders from time-to-time so that they can work upon them and improve accordingly.

3.4 ICICI Lombard General Insurance is a major general insurance company in India. It utilises a plethora of data from alternative sources which gives it an edge in different processes like risk underwriting, claims settlement and customer acquisition. Some of the alternative data sources used by them are mentioned below:

a. **Wearable devices:** ICICI Lombard captures information from wearable devices like fitness bands and smartwatches which provide insights into the physical activities, sleep schedules and BMI index of policyholders. The data collected from these devices provides valuable insights into a policyholder's lifestyle and overall health habits.
b. **Face scan:** ICICI Lombard has integrate a face scan feature into its app that shows vitals such as blood pressure, oxygen, heart rate, respiration rate and stress level within minutes. The underlying technology used by the app is photoplethysmography. [5]
c. **Telemedicine platforms:** ICICI Lombard may forge associations with telemedicine platforms that offer online doctor consultations and other healthcare services. By integrating these platforms, ICICI Lombard can access information about policyholders' virtual doctor visits, prescribed medications and treatment plans. This leads to more comprehensive understanding of the insured persons' behaviours and better risk assessment.
d. **Use of IoT devices to cut risks:**[6] IoT devices are being used to detect the possibility of risks earlier and mitigate the same with timely actions. This would prevent any loss to the policyholder and eventually less claims burden on the insurer.
e. **Underwriting via chatbots:**[7] Chatbots have the capability to accurately serve requests from multiple customers simultaneously with quick turnaround time. They capture the data points from the customer and generate a customised quotation on the spot. This also helps in increasing customer satisfaction and better sales conversion ratio.

How does it help ICICI Lombard Insurance?

a. **Risk assessment:** The personal data of policyholders can provide better understanding of their risk profile, leading to better underwriting and risk assessment. This can further lead to reduced pricing for customers with active and healthy lifestyles.
b. **Wellness programmes:** ICICI Lombard may offer customised wellness programmes to policyholders, incentivising them to take up habits like regular exercising, taking a

[5] https://timesofindia.indiatimes.com/business/india-business/icici-lombard-app-now-incorporates-face-scan-diagnostics-features/articleshow/89323759.cms
[6] https://timesofindia.indiatimes.com/business/india-business/icici-lombard-uses-iot-to-cut-risks-in-commercial-covers/articleshow/89601208.cms
[7] https://www.icicilombard.com/experts-blogs/story/when-technology-meets-insurance

balanced diet, etc. By promoting wellness, ICICI Lombard aims to improve policyholders' overall health outcomes and reduce the likelihood of claims.

c. **Claims management:** In the event of a claim, the health data collected from wearable devices and fitness apps can play a crucial role in claims management. This data can provide valuable insights into the policyholder's health condition before and after an incident, helping ICICI Lombard validate claims and expedite the claims settlement process.

By leveraging alternative data sources in health insurance, ICICI Lombard aims to offer more tailored coverage, promote healthier lifestyles among policyholders, improve risk assessment accuracy, and enhance the overall insurance experience for their customers.

4. Different data types available through alternative data sources

Alternative data is of various types – structured, semi-structured or unstructured. Data points also come in a batch mode or real-time streaming mode. The type and frequency are critical to understand the mechanism of ingestion and data processing for obtaining desired and sustainable business outcomes. The various formats are explained below:

4.1 Structured data:

Structured data is organised data with fixed fields and columns such as date, address and age. This data is easily understood by the machine language. Also, the most attractive feature of this data type is that it can be extracted very easily from relational databases.

4.2 Unstructured data:

Unstructured data is the opposite of structured data and is highly complex and unorganised. As it does not conform to a fixed standard, it's not possible to store this type of data in relational databases. It's also referred to as big data. Some examples of unstructured data are social media and multimedia data such as images, audio and video. Unstructured data needs to be parsed to extract structured information using various types of tools or algorithms like optical character recognition (OCR) and natural language processing (NLP). Post that, it can be integrated with traditional data for analysis.

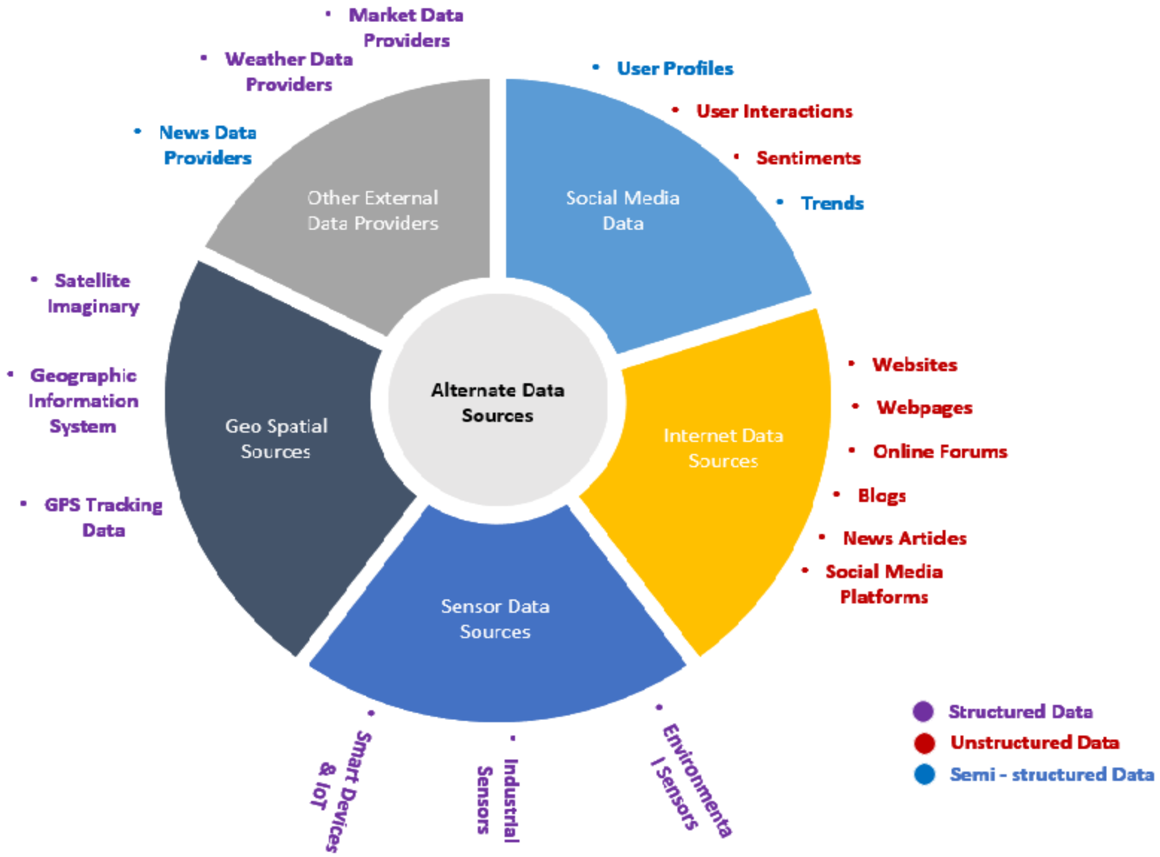
4.3 Semi-structured data:

Semi-structured data lies between structured and unstructured data. It has some organisational structure or tags but does not fit into a rigid schema. Semi-structured data may have a variable number of attributes or fields, allowing for flexibility in the data model. Examples of semi-structured data include XML files, JSON files, HTML documents and log files.

4.4 Streaming data:

Streaming data is generated continuously and in real time. It refers to data that is transmitted or received in a continuous flow rather than being stored in batches. Streaming data is commonly produced by sensors, IoT devices, social media platforms and financial markets. Analysing streaming data requires specialised tools and techniques for real-time processing and analysis.

5. Different types of alternative data sources



Types of alternative data sources

a. **Social media data:** Social media platforms generate vast amounts of data through user profiles, interactions, sentiments and trends. This data can be used for market research, sentiment analysis and customer insights.

Some of the common social media platforms are:

1. **Instagram data:** Analysing posts to understand broader level user sentiment and get insights related to social and economic stability of a group or region
2. **Facebook data:** Analysing Facebook data for customised underwriting to check if the individual is involved in high-risk activities, and analyse his/her medical history and general habits

Internet data sources: Internet data includes information from websites, web pages, online forums and blogs. It can be scraped, aggregated and analysed to extract valuable insights and trends.

Some of the main sources of this data are follows:

1. **E-commerce data:** Analysing data of e-commerce websites and applications like transactions, purchase history etc., can provide significant insights into customer behaviour which can be used for cross-selling and understanding fraudulent activities.
2. **Mobile app data:** Data from various mobile apps like HealthifyMe and Garmin can be analysed to get insights into customer behaviour to check if he/she is following a healthy diet regime, working out adequately or has a sedentary life.
3. **Online reviews and ratings:** Customer reviews and posts online can provide very useful insights such as current market trends, receptivity of the product, and the likelihood of them recommending it to someone.

c. **Sensor data sources:** Sensors can be of any type – heat, noise, sound, etc. This data can be used to capture and transmit information from the physical world in real time. Sensors can be broadly classified into environmental sensors (room temperature, humidity) and industrial sensors (capture and monitor performance of various processes and equipment).

Some of the sensor data sources are explained below:

1. **Telematics data:** Extracting data from vehicle sensors and GPS maps to derive patterns of vehicle movement, driving behaviour etc.
2. **IoT device data:** Extracting data from IoT devices installed in properties to get real-time information of fire, flood, earthquake and theft risks for loss prevention and seamless claims settlements
3. **Wearable device data:** Extracting data from wearable devices like fitness bands and smartwatches to assess a person's lifestyle for accurate and customised pricing and underwriting

d. **Geospatial data:** These kind of data sources provide geographic information like GPS, GIS and satellite imagery data on a real-time basis to enable detection of location-based risks and analysis of spatial relationships.

Some geospatial data sources are as follows:

1. **Weather data:** Incorporating weather information to assess risks related to natural disasters, property damage and claims management
2. **Satellite imagery:** Analysing satellite images to assess property conditions, vegetation health or detect fraudulent claims

e. **External data providers:** These are third-party sources that provide specialised data services. Examples include market data providers, credit rating data providers and news data providers. Organisations often integrate these external data sources to augment their existing datasets.

Some types of external data providers are discussed below:

1. **Vehicle registration data:** Accessing vehicle registration databases to verify vehicle information, ownership history or identify potential fraud
2. **Health data:** Fetching health data from electronic health records or health monitoring apps/websites to understand medical history of the policyholders.
3. **Travel data:** Analysing travel data from booking platforms, travel agencies, online travel websites etc., to understand travel patterns of policyholders and assessing risk accordingly
4. **Environmental data:** Analysing environmental data like pollution levels and proximity to flood-prone areas to assess risks associated with a particular property
5. **Vehicle repair and maintenance data:** Analysing vehicle service and claims data to understand claims history, frequency of service etc., to evaluate vehicle condition better and set price accordingly
6. **Demographic data:** Analysing demographic data such as population density, income levels etc., from public records and surveys to assess cross-selling opportunities and carry out the product design
7. **News and media data:** Analysing news articles, social media posts and blogs to identify emerging trends, customer behaviour and risk factors
8. **Credit card transaction data:** Analysing credit card transactions data to understand customers' spending behaviour and provide customised insurance offerings

[1]

6. Challenges of using alternative data sources

As we have seen, alternative data is becoming an integral part of the insurance value chain. In times to come, the adoption of alternative data by insurance companies is only going to increase. However, the usage of alternative data does pose certain challenges as one needs to find the 'right' data that not only supports the insurer's use cases but is also compliant with the laws and regulations of that geography.

Some of the common challenges associated with data from alternative data sources are:

1. **Incomplete or unverifiable data:**

   Not all data sources provide complete or verified data. Incomplete data can prevent the system from producing effective statistical estimates, and unverifiable data cannot be trusted even if it leads to some seemingly valid estimates or results.

2. **Regulatory and compliance:**

   One of the major challenges with alternative data sources is ensuring regulatory compliance. Different industries and regions have specific regulations and guidelines governing the collection, storage and use of data. When using alternative data sources, organisations need to ensure that they adhere to these regulations to maintain compliance.

3. **Large amount of data and non-standard format:**

   Data from alternative data sources is often unstructured and large in volume because of which it becomes very difficult to process it in a continuous and real-time manner over prolonged periods.

4. **Data relevance:**

   Although there are various alternative data sources available today, many of them are not relevant for insurance-specific use cases. Therefore, filtering out the non-relevant data can be a challenging and ambiguous exercise for the insurance player.

5. **Regulatory and ethical concerns:**

   Although data from alternative data sources can be very useful for insurance players' use cases, there can be regulatory and ethical concerns with the same data. It can be because of the manner, in which the data was procured or was intended to be used or because of any bias in the usage by the downstream user. These concerns must be carefully and transparently addressed to prevent any future legal consequences.

6. **Credibility:**

   Establishing the credibility of the alternative data source is a major challenge as there is little transparency provided when it comes to critical questions like underlying source of the data, data quality checks implemented, predictive power of data, etc.

7. **Lack of data archives:**

   Data sources often have limited or no archives or historical data required for back-testing the effectiveness of data usage over a long period of time.

To address these challenges, a comprehensive approach involving legal and regulatory compliance, technical infrastructure, data management processes, ethical frameworks and partnerships with reliable data providers is needed. Organisations should establish clear guidelines and practices for the selection, evaluation, integration and analysis of alternative data sources to maximise their benefits while mitigating associated risks.

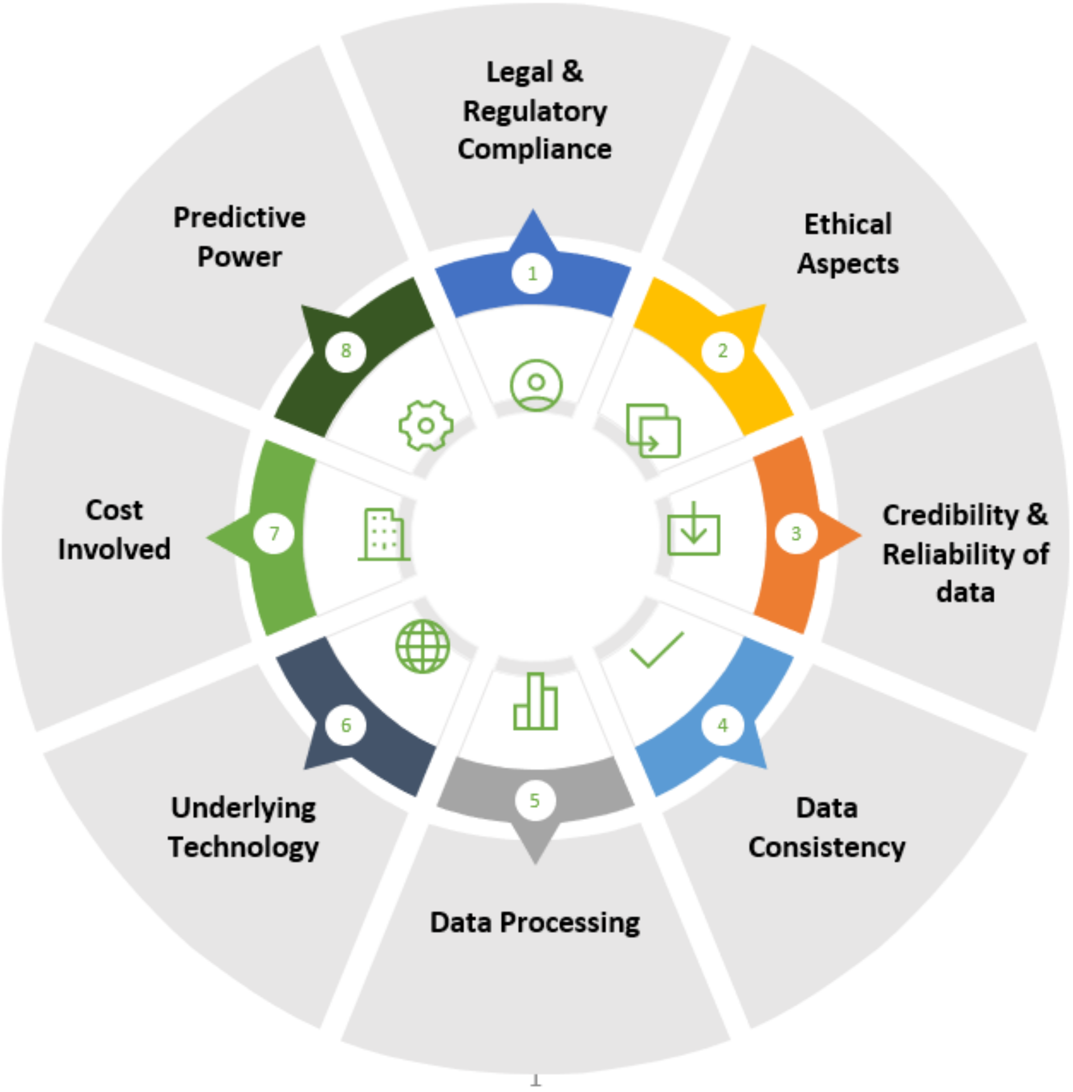
7. Problem statement

Evaluating an alternative data source for an insurance company can be a challenging task as it involves issues related to quality, recency, ability to predict or provide desired outcomes, accuracy, etc.

To verify the credibility of a particular data source, it's imperative for insurers to adopt a comprehensive framework that enables them to determine whether the source would meet the above stated goals.

8. Solution – a framework to evaluate alternative data sources

The framework defines key parameters to identify alternative data sources and evaluate them. These key parameters are shown below.



Key evaluation parameters

a. Legal and regulatory compliance

Data source is a legal entity and complies with all applicable local and international laws. The following guidelines can be checked for legal and regulatory compliance of data.

Sr. no.	Guideline	Risk level	Will it impact the alternative data source?
1	Check if data sources are compliant with local and international laws, regulations, and policies.	Low – non-compliant with applicable local and international laws	Yes
		Medium – compliance in place with the most applicable laws, with all non-compliance issues identified and addressed	
		High – compliance with all applicable laws and regulations.	
2	Check whether proper data governance practices are in place to ensure that the data has been collected, processed and analysed in a compliant and secure manner.	Low – inadequate data governance practices requiring urgent intervention	Yes
		Medium – adequate data governance practices with minor issues to be resolved	
		High – sufficient data governance practices (data lineage, data quality) and no need for urgent intervention	
3	Check if there are procedures in place for end users to file complaints in case of any issues regarding the collection and usage of data.	Low – non-existent procedures	Yes
		Medium – existing procedures with scope for improvement	
		High – sufficient procedures for end users to file complaints and get their grievances addressed	
4	Confirm if the alternative data sources are subjected to regular audits by professional third parties.	Low – no provision of auditing	Yes
		Medium – auditing provisions with scope for improvement	
		High – sufficient audit provisions	
5	Check whether data sources have sufficient security measures in place to prevent any unauthorised access to data, malware attacks etc.	Low – required security measures absent	No
		Medium – required security measures in place with some scope for improvement	
		High – required security measures present	
6	The alternative data source is subjected to data retention as per the regulatory requirements.	Low – temporary data retention; does not comply with regulatory requirements	No
		Medium – data retention as per regulatory requirements with scope for improvement	
		High – data retention as per regulatory requirements with the flexibility to enhance	

		the duration of storage as per the requirement	
7	Check is audit trails are maintained.	Low – audit trails not maintained	No
		Medium – basic audit trails maintained to demonstrate compliance with regulatory requirements	
		High – comprehensive audit trails maintained	
8	The data source relies on third-party relationships, such as vendors or contractors. Evaluate whether there are controls in place to ensure that these third parties adhere to the same accountability standards as the data source.	Low – no controls in place to ensure that third parties adhere to the same accountability standards as the data source; third parties not held accountable	No
		Medium – some controls to ensure that third parties adhere to the same accountability standards as the data source, while having areas that require improvement or clarification	
		High – controls in place to ensure that third parties adhere to the same accountability standards as the data source, with clear contractual agreements outlining accountability responsibilities	
9	Restrictions are present from the source owner to retain data at the final consumption stage (allowable within regulations).	Low – restricted	No
		Medium – partially restricted	
		High – no restriction	
10	Restrictions are present from the source owner for using the data in a specific way.	Low – restricted	No
		Medium – partially restricted	
		High – no restriction	

b. Ethical aspects

Data is sourced and processed in an ethical manner with full transparency and acknowledgment of the users.
Following guidelines can be followed to check ethical aspect of data:

Sr. no.	Guideline	Risk level	Will it impact the alternative data source?	Rating (to be filled in by the user)
1	Data has been sourced responsibly, without exploitation or manipulation of the users.	Yes – at least one user has been exploited	Yes	
		No – no users exploited		
2	The data source is being used for an ethical, intended purpose.	Yes – data is not used for the intended ethical purpose at time of collection	Yes	
		No – data used for the ethical purpose for which it was intended at time of collection		
3	Consumption of the data is free from any bias and is used in a manner that is fair to all individuals	Yes – consumption of the data source yields biased results	Yes	
		No – consumption of the data source does not yield any bias		

c. Credibility and reliability of data

Data should be fetched from credible and reliable data sources that can be backed using proven research. Moreover, sources should be unbiased and certified or authentic. Following guidelines can be followed to check the credibility and reliability of data:

Sr. no.	Guideline	Risk level	Will it impact the alternative data source?	Rating (to be filled in by the user)
1	Check the source – is the author, publisher or sponsor of the data reliable/ credible?	Low – source is unknown or unverified	No	
		Medium – source is known but unverified		

		High – source is known and verified		
2	Check for any potential bias in the data – e.g. selection bias, confirmation bias or cultural bias.	Low – data has bias that can materially impact the purpose for which it will be used	No	
		Medium – data has bias but the impact is less material or can be adjusted		
		High – data doesn’t appear to have bias, or is fit for purpose		
3	Ensure source data security.	Low – non-secure data, stored in a shared environment where it can be altered easily	No	
		Medium – data is stored in a shared environment, with limited people having editable rights		
		High – data is stored in a separate environment with editable rights given to only limited people		
4	Ensure relevance of data.	Low – source and data not relevant for research	No	
		Medium – source and data partially relevant for research		
		High – source and data relevant for research		

d. **Data consistency**

Sourcing of data should be consistent over different periods to facilitate easy processing and avoid any discrepancies in the overall process.

The following guidelines should be checked to ensure the consistency of data:

Sr. no.	Guideline	Risk level	Will it impact the alternative data source?	Rating (to be filled by the user)
1	Check for consistent data point availability.	Low – frequently varying availability of data points (data set should be clean and	No	

		null data points should not be high)		
		Medium – data points available for only half of the time		
		High – 90% availability of data points		
2	Data should have a well-defined data dictionary.	Low – data does not have a data dictionary	No	
		Medium – dataset has a partially written data dictionary		
		High – data dictionary is well defined for every attribute, including information of primary keys and list of values allowed per attribute		
3	Data format should be consistent.	Low – inconsistent data format	No	
		Medium – partially consistent data format		
		High – consistent data format for more than 90% of the data		
4	Data should have a unique identifier.	Low – no unique identifier	No	
		Medium – logic required to identify a unique identifier per record		
		High – unique identifier present in data		
5	Ensure availability and implementation of data quality checks.	Low – quality checks not available and not implemented	No	
		Medium – checks available but partially implemented		
		High – checks available and fully implemented		

e. Data processing

Data should be seamlessly and quickly available from the source as and when it's required by the downstream applications.

Following guidelines can be followed to check the pace of data processing:

Sr. no.	Guideline	Risk level	Will it impact the alternative data source?	Rating (to be filled in by the user)
1	Check if the format of the data is well structured and in an analysable format (e.g. .csv, .JSON).	Low – unstructured or proprietary data format	No	
		Medium – structured data format but not available in standardised format; less effort required for parsing		
		High – data format is structured and is available in a standardised format; no effort required for parsing		
2	Check if data can easily be ingested from the source.	Low – cost and effort of integration between data sources and ingestion tools is either expensive or with considerable practical challenges	No	
		Medium – cost and effort of integration between data sources and ingestion tools is in an acceptable range		
		High – cost and effort of integration between data sources and the ingestion tools is minimal		
3	Check if the data volume provided by the source can be easily processed by automation tools.	Low – data volume is too high or too low	No	
		Medium – data volume can be managed with some changes		
		High – data volumes can easily be processed		
4	Check whether missing	Low – data transformation required for more than 50% of data	No	

	data/null handling is required.	Medium – data transformation required for 30–50% of data		
		High – data transformation required for less than 30% of data		
5	Ensure ease of use.	Low – considerable runtime and manual specialist interpretation needed to obtain required information/statistics for a given schema	No	
		Medium – considerable runtime and some manual intervention needed to obtain the required information/statistics for a given schema.		
		High – simple and automated processes to obtain required information/statistics for a given schema.		
6	Check process outcomes.	Low – extracted information/statistics cannot fit into internal processes; no indicators available for impact quantification	No	
		Medium – extracted information/statistics do not fit readily into internal processes; only proxy indicators available for impact quantification		
		High – extracted information/statistics align with internal processes; additional impact quantifiable		

f. Underlying technology

The underlying technology assessment is important to understand if the source tech stack is compatible with the insurer's or end user's tech stack.

Following guidelines can be followed to check the recency of data:

Sr. no.	Guideline	Risk level	Will it impact the alternative data source?	Rating (to be filled in by the user)
1	Confirm whether the technology used is in existence and widely used across the industry.	Low – relatively old and obsolete technology	No	
		Medium – relatively old but mature technology		
		High – widely and commonly used technology across the industry		
2	Is the source black-box, grey-box or white-box implementation?	Low – black box; components not exposed, output untraceable	No	
		Medium – grey box; some components exposed and traceable		
		High – white box; all components exposed and traceable		
3	Check if the source provider customises data as per different requirements.	Low – no customisation possible	No	
		Medium – partial customisation available		
		High – high degree of flexibility and customisation possible		
4	Check deployment options.	Low – either on-prem or cloud option available as infrastructure as a service (IaaS)	No	
		Medium – both on-prem and cloud options available as IaaS		
		High – both on-prem and cloud available as software as a service (SaaS) and available in the marketplace		
5	Evaluate the scalability of accessing and processing alternative	Low – identifying potential limitations in accessing and processing data from	No	

	data sources. Consider whether the data sources can accommodate the scale and volume of data needed for your insurance operations. Assess if the data acquisition and integration methods can handle increased data loads.	alternative sources at large scales		
		Medium – evaluating the potential for data sources to scale and accommodate growing data needs		
		High – ensuring that the chosen alternative data sources have the necessary infrastructure and capacity to handle significant increases in data volume		
6	Check if the data source refreshes at the required frequency for a defined use case.	Low – latest data not available	No	
		Medium – latest data available till a pre-defined period to get envisaged business outcomes		
		High – latest data available – sometimes even on daily/real-time basis		

g. Cost involved

The costs involved in acquiring data from sources should be reasonable to ensure continuous and long-term processing of data.

Following guidelines can be followed to check the cost involved:

Sr. no.	Guideline	Risk level	Will it impact the alternative data source?	Rating (to be filled in by the user)
1	Is the data available for free?	Low – paid vendor-sourced data	No	
		Medium – paid government-sourced data		
		High – free data		
2	Check the cost for a commercial model.	Low – one-time payment	No	
		Medium – small one-time payment with a recurring pay-as-you-go (PAYG) model		
		High – PAYG model or no upfront payment		
3	Check the acquisition cost of data versus its return or impact.	Low – high acquisition cost as compared to the envisaged return or impact	No	
		Medium - medium acquisition cost as compared		

		to the envisaged return or impact		
		High – low acquisition cost as compared to the envisaged return or impact		
4	Check the cost of automating and processing the data.	Low – high cost of processing the data	No	
		Medium – medium cost of processing the data		
		High – low cost of processing the data		

h. Predictive ability

Following guidelines can be followed to check predictive power of data:

Sr. no.	Guideline	Risk level	Will it impact the alternative data source?
1	Ensure sufficiency of data volume for each important variable.	Low – low volume	No
		Medium – moderate volume	
		High – high/sufficient volume	
2	Check whether the ability to deal with abnormalities in data is present.	Low – ability missing, outlier values present	No
		Medium – moderate ability; outlier values present but can be treated to improve the accuracy of the model	
		High – ability present; outlier values absent or very few in number which can be treated easily	
3	Check if the data has a sufficient number of variables to improve accuracy.	Low – very few variables, variables that are too linearly dependent, or too many variables leading to overfitting	No
		Medium – enough variables to provide a reasonable level of prediction	
		High – ideal number and types of variables, providing sufficient prediction accuracy	
4	Check the source of data. Data generated through automation basis an activity is more reliable than data that is manually inputted	Low – high level of manual inputs for one or more variables	No
		Medium – moderate level of reliance on manually inputted data	
		High – high level of automated data generated from an activity	

	(e.g. consumption of water in a day).		
5	Check the period for which the data is available (considering end user/use cases).	Low – data available for short periods; periods of long, unusual events which make the data stale	No
		Medium – data available for a reasonable period; no unusual events	
		High – data available for the right period of time; no unusual events	

9. Sample data source evaluation as per the framework

As discussed in the previous sections, we now know what parameters can be considered while evaluating an alternative data source and whether we should use the same as a trusted and reliable source of information for insurer's day-to-day processes. Now, we will evaluate some of these data sources using the same framework to see how this works in real time.

Below is a sample report which was generated at the end of our evaluation using this framework.

Sr. no.	Assessment parameter	Overall rating	Comments
1.	Legal and regulatory compliance	Low	Avoid using data for legal/compliance-related use cases.
2.	Ethical aspects	High	As long as the source is verified and purpose of data usage is defined, we can use the framework.
3.	Credibility and reliability of data	Medium	Use for activities like customer acquisition and marketing and avoid for legal or compliance-related use cases.
4.	Data consistency	High	Data is consistent and can be used for all use cases.
5.	Data processing	Low	It'll be difficult to automate the data extraction from its source. Consider opting for one-time or batch processing.
6.	Cost involved	High	Cost involved is reasonable.
7.	Underlying technology	Medium	Insurers may need to customise their systems, depending on the data source.
8.	Predictive ability	High	Use for predicting trends, pricing etc.

a) Case study 1: Using physical activity of potential customers to predict mortality/morbidity rates

1. Line of business: Life and health insurance

2. Data sources

- Human API
- Terra API
  However, Terra API requires users' consent to pull data.

3. Implementation examples

Here, we are highlighting some entities that have implemented the mentioned use cases in their processes.

Vitality: John Hancock in the US – focused on diabetic patients
Vitality: AIA in Hong Kong – 10% extra cover + rewards
Vitality: Discovery in South Africa – discounts on healthcare and non-healthcare items

4. Impact on value chain

- **Primary:** Underwriting
- **Secondary:** Marketing – running campaigns on upselling and customer retention

5. Does it augment or replace traditional data?

When combined with existing clinical factors, like blood pressure, rather than replacing them, it improves our understanding of the underwriting risk.

6. Assessment parameters

- **Legal and regulatory compliance**
  There are no legal or regulatory concerns. Insurers across the globe already use activity data to offer benefits to policyholders. For example, Vitality works with insurers in the US, South Africa and Hong Kong. Similarly, there have been instances of health insurers in India offering rewards based on physical activity.
  Rating: **High**

- **Ethical**
  There are concerns about the unhealthy population being at a disadvantage while using physical activity data to predict mortality rates as they might be charged considerably more or denied insurance altogether.
  Rating: **Medium**

- **Credibility and reliability of data**
  Barring step count, the accuracy of every other parameter is less reliable.
  Rating: **Low**

- **Data consistency**
  Data is consistently available from the sources with defined parameters. Rating: **High**

- **Data processing**
  There are many service providers that have built an API wrapper on top of source data from hundreds of devices.
  - Human API
  - Terra API

  Rating: **High**

- **Underlying technology**
  The technology is fairly evolved and widely used.
  Rating: **High**

- **Cost involved**

  - **One-time cost:** The cost of smartwatches starts from INR 5,000 in India, which is generally borne by the policyholder. Insurers might thus be required to invest in systems to fetch information from these watches by calling APIs.
  - **Ongoing costs:** Although the exact costs are not known, most watch manufacturers extend APIs to export data, which is likely to be inexpensive or free of charge.

  Rating: **High**

- **Predictive ability**

  - **Presence of a paradox** – occupational vs leisure time activity
    For example, even though manual labourers can average 20,000 steps a day – which would ideally indicate good health and fitness – in reality, such extended activity may cause sustained inflammation and increase the 24-hour heart rate and blood pressure. If prolonged, it can also impair cardiovascular health and increase mortality risk.
  - **Not all activities can be treated equally**
    Even when different physical activities use equivalent amounts of energy (METs), they don't always offer the same health benefits. For example, activities that require dynamic use of large muscle mass, like swimming or racquet sports, are associated with lower all-cause mortality and cardiovascular disease (CVD) risk compared to sports that consume similar METs but don't use as many different muscle groups, such as running.
  - **Multiple tracking metrics**
    Steps and distance are more commonly recorded but heart rate, duration and frequency are more accurate.

  Rating: **Medium**

b) Case study 2: Enhancing the property insurance value chain using alternative data sources

1. Line of business: Property insurance

2. Data sources

- IoT sensors installed in buildings to monitor temperature, humidity and security where data could be taken from various sensors in the buildings
- Social media data to assess the reputation and behaviour of homeowners or businesses

3. Implementation examples

- Insurers utilise IoT sensors in commercial buildings to detect potential risks, such as water leaks or fire hazards, and provide proactive risk mitigation measures.
- Social media data is analysed to identify potential fraud or misrepresentation in property insurance claims.
- Satellite imagery is utilised to assess property damage after natural disasters and expedite claims processing.

4. Impact on value chain

- **Primary:** Underwriting – assessing risks and determining premiums based on real-time data from IoT sensors and satellite imagery
- **Secondary**: Claims management – expediting claims processing and verifying damages using satellite imagery, IoT sensors and public records

5. Does it augment or replace traditional data?

This augments traditional data. Alternative data sources enhance the understanding of property risks and enable more accurate underwriting and claims management processes.

6. Assessment parameters

- **Legal and regulatory compliance**
  Ensure that extracting data is compliant with the local laws and that we have the consent of the users.

Data source	Rating
IoT devices	High
Instagram	Low

- **Ethical aspect**
  Obtaining necessary consent from policyholders or property owners for the collection and usage of IoT data will be easy. However, it might be difficult to do the same for social media platforms.

Data source	Rating
IoT devices	High
Instagram	Low

- **Credibility and reliability of data**
  Data from IoT devices will be highly reliable as we have calibrated and tested these devices before installation. However, data from social media channels cannot be termed as highly reliable because of intermediaries processing the same.

Data source	Rating
IoT devices	High
Instagram	Medium

- **Data consistency**
  IoT devices will share highly consistent data. However, data obtained from social media channels won't be as consistent as it would vary according to consistent usage, intermediaries' operations etc.

Data source	Rating
IoT devices	High
Instagram	Low

- **Data processing**
  Ease of automation will be high.

Data source	Rating
IoT devices	High
Instagram	High

- **Underlying technology**
  The underlying technology is relatively new.

Data source	Rating
IoT devices	Medium
Instagram	Medium

- **Cost involved**

For IoT devices, there will be a single one-time cost involved. For other sources, like social media platforms, continuous costs of data acquisition will be present.

Data source	Rating
IoT devices	High
Instagram	Low

- **Predictive ability**
  - Risk detection and mitigation: IoT sensors installed in buildings can provide real-time data on various parameters such as temperature, humidity or security breaches. By incorporating this data into risk models, insurers can enhance their ability to detect and mitigate risks.
  - Social media data gives an overall depiction of the group or society. However, it’s not as accurate as the data from IoT devices.

Data source	Rating
IoT devices	High
Instagram	Medium

Appendix A: Members of the Joint Working Party

To better leverage alternative sources of data leading to a better quantification of risks at a more granular level, working party was constituted with the following members on 4th March 2023.

1. Sumit Ramani – Chairperson
2. Abhijit Pal
3. Anjani Choudhary
4. Amit Tiwari
5. Devadeep Gupta
6. K S Gopalakrishnan
7. Karan Vashisht
8. Raghavendra Pawar
9. Sunil Padasala
10. T Balachandra Joshi
11. Hetal Shah