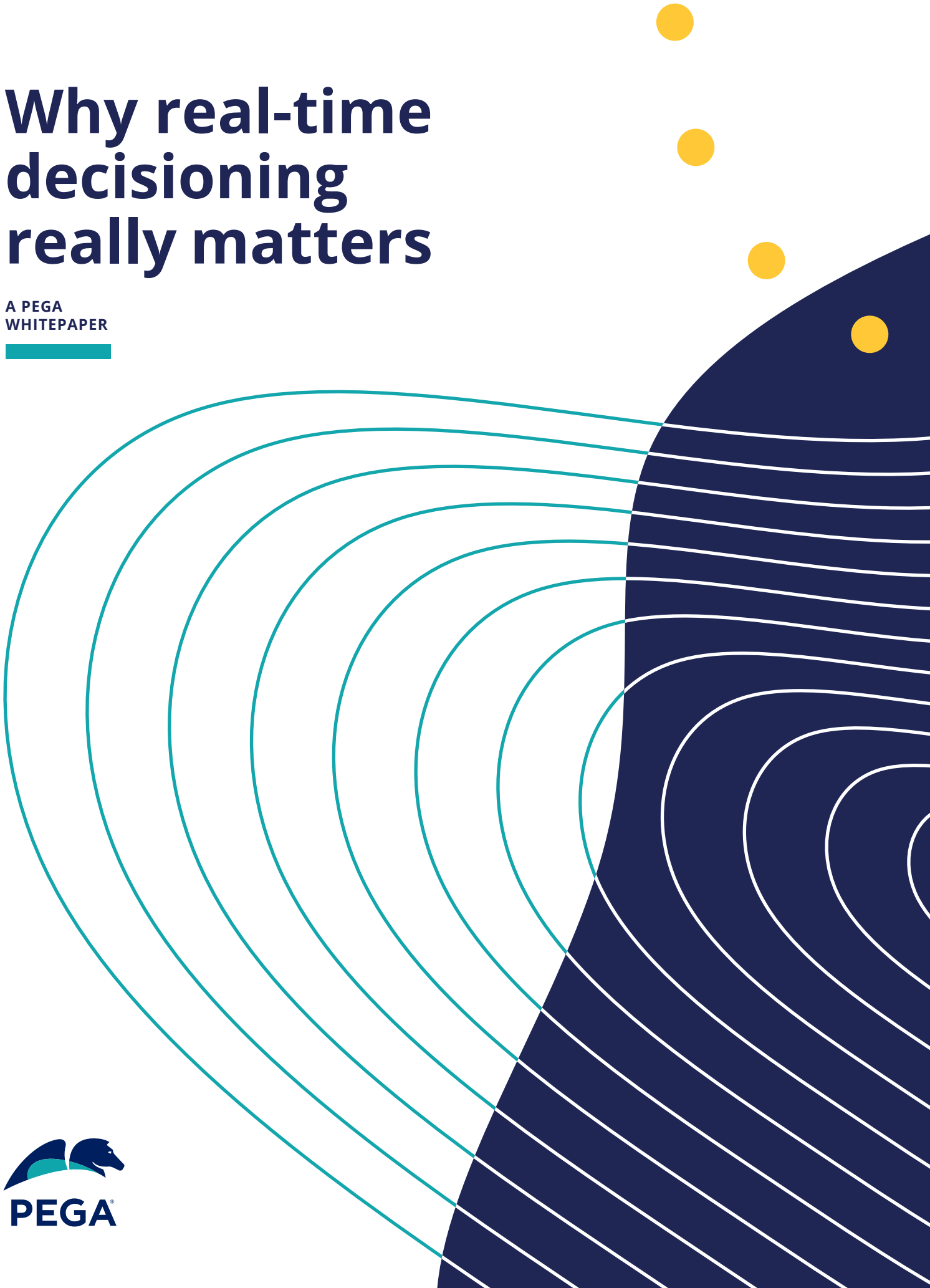
Why real-time decisioning really matters

A PEGA
WHITEPAPER



PEGA

Introduction

Customer experience (CX) has hit a wall across many industries. Companies are using outdated approaches and technologies to try and solve complex engagement problems – and they're failing. This is generating a huge amount of frustration:

- Consumers are frustrated because their interactions with businesses are often awkward, tedious, and irrelevant. They'll have a great experience, followed by 10 throw-away moments. They feel like they're constantly getting a sales pitch.
- Business leaders are frustrated because they've invested in solutions that should have "righted the ship" and set them apart from the competition, but have only produced sporadic improvements with no significant return on investment.
- Investors are frustrated because businesses have been paying lip service to customer experience – but complaints are still going viral. Market share is evaporating, and dividends are declining.

The underlying issue: Business needs versus customer problems

The real challenge isn't that consumer expectations are too high or that businesses are reluctant to invest in CX – neither is true. Consumers simply want more of the "best" experiences, like when companies anticipate their needs, make information easy to find, and seem to truly appreciate their business. Likewise, firms have spent untold billions on customer experience to meet and exceed customer expectations. But they've seen very little in return, after buying into promises made by vendors, agencies, and partners. However, firms keep spending because they're still losing ground.

The real problem is that these solutions were built to support business needs, not solve customer problems.

Marketing automation suites, service desktops, model-building tools, and business intelligence (BI) platforms are part of an entire ecosystem of solutions designed to make it easier for a business to move offers and messages into the market. Those solutions gather data from disparate systems in batches, consolidate history, analyze behaviors, append third-party information, create rules, build segments, optimize targeting, and generate campaigns. They enable businesses to engage customers at an exponentially faster rate, and at a tiny fraction of the previous cost. However, they ignore one critical issue:

People are complex creatures that act on impulse and change their minds quickly. When they consume new info, it often sends them down a new path seconds later.

People are fluid, dynamic, and adaptable. That's why an ecosystem built around batched data, relational databases, segment-based decisions, and campaign-based engagement can't satisfy modern customer experience requirements. These technologies were built to pick a message then push it to large audiences or segments at scale – not to analyze an individual customer's needs at a specific moment and quickly address them before they change.

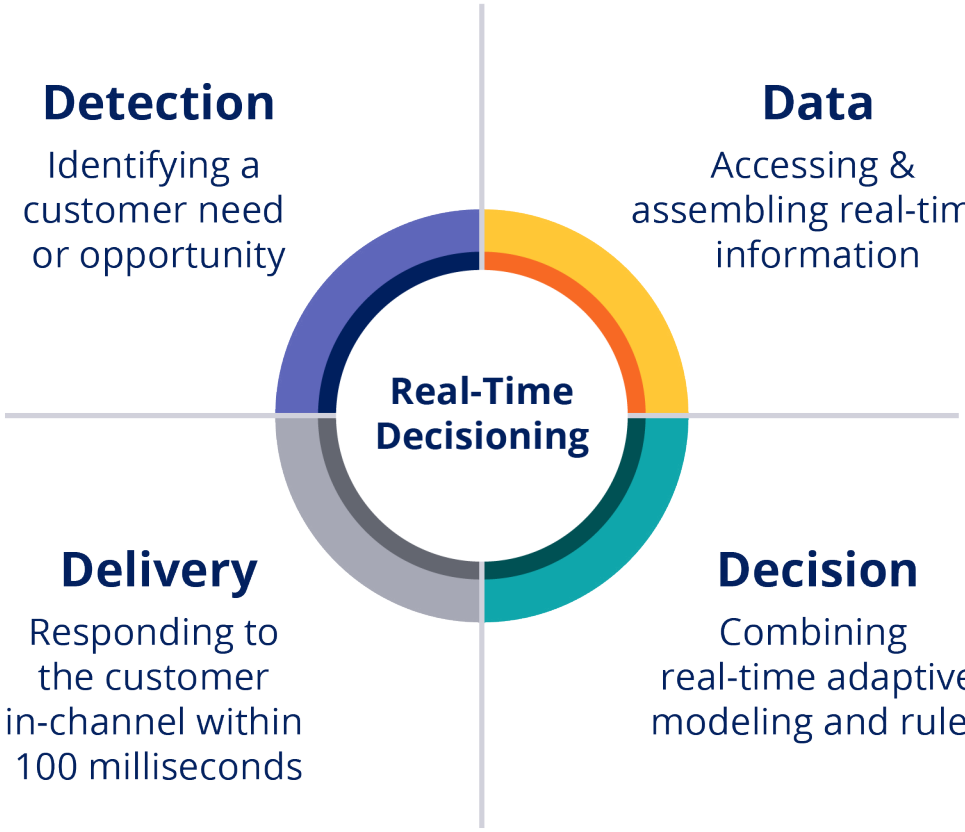
The competitive advantage of real-time decisioning

Your customer's context (their situation, environment, emotions, motivations, and behaviors) is constantly changing as they consume information. Moments of opportunity will open, then close, within seconds. But those same individuals are exposed to a massive number of competitive messages each day, each of which looks and sounds very similar to your own. You won't stand out if you can't read the signals and optimize your message within that narrow window. You won't lose the opportunity because your competitors were better, necessarily. You simply couldn't move fast enough, and the competition was there – in the right place, at the right time. As a result, real-time decisioning is now directly linked to having the competitive advantage.

Brands that can "re-decision" a customer five, 10, 20, or even 50 times within a single experience will take the lead. Each time they activate new data and leverage predictive models to re-assess a customer's propensity to take action, they become far more relevant than brands who cannot – gaining an **enormous competitive advantage** in the process. This paper will explain how those rapid, real-time decisions are made, what data and analytics are used to power them, and how quickly an organization truly has to move to make a significant impact on the customer experience.

The four cornerstone capabilities of real-time decisioning

Enabling relevant and contextual interactions involves more than just the real-time decision. In total, there are four factors that go into an optimized, real-time decisioning framework: detection, data, decision, and delivery. Understanding these will help you assess your own real-time readiness.



Detection

Identifying a customer opportunity

Detecting an opportunity is the primary building block of a real-time engagement program and the first aspect that must be assessed. Simply put, none of the downstream abilities or processes will matter unless you're aware of the opportunity to sell, serve, or retain.

Utilizing preemptive detection

The most important aspect of detection is that it can't be scheduled. In today's market, it must be sensed by analyzing contextual data. Digital windows of opportunity open and close in seconds. This is a stark contrast from years past, when it was common practice to schedule batch processes to run on a nightly, or even weekly basis.

Your ability to sense these moments of opportunity as they happen, or even before, will require you to enable event detection that uses both fast- and slow-moving data to predict what's likely to happen. Many use cases will require near-instantaneous detection – like mining high-speed data streams for rare events, especially ones with a small window of opportunity. Sometimes simple event triggers (ie. if this happens, then do that) may be enough to do the job. But the most powerful emerging use cases are built on rugged landscapes containing complex, interconnected events where more sophisticated capabilities are required.

The most important aspect of detection is that it can't be scheduled. In today's market, it must be sensed by analyzing contextual data.

What are the different event types?

Complex event detection searches for patterns across data sources and seeks to identify when a potential moment of opportunity may exist.

- Simple events: Listening to a single data source (e.g., click on a link)
- Non-events: Event that was expected but did not happen (e.g. regular bank deposit, mobile or online login, branch visit, etc.)
- Complex events: A combination of multiple simple events and non-events that together represent a potentially important moment

Detecting your window of opportunity

Detection is a crucial building block of real-time decisioning. If your capacity for detection isn't fast enough, you'll likely miss your window of opportunity entirely. This makes everything else (data assembly, decision-making, and delivery) irrelevant – as you've already missed your chance.

Scenario 1: Detection happens too slowly, leaving no time to react.

Scenario 2: Detection happens quickly, allowing sufficient time to react.

Illustration: Latency of detection and the window of opportunity

The design of digital experiences often requires response times of 100 – 200 milliseconds (ms).

How fast is fast enough?

Your business' requirement for detection speed – and for end-to-end execution – may be measured in microseconds (one millionth of a second) for some use cases, such as online trading and real-time ad buying. Digital experiences, such as web clicks and mobile app interactions, require response times of less than half a second. And the design of digital experiences often require response times of 100 – 200 milliseconds (ms). Time frames beyond one second can lead to progressively sub-optimal or even detrimental outcomes.

The telecommunications industry is relatively advanced in its approach to real-time decisions. For example, a consumer topping off credit on a prepaid phone can become an opportunity to present a contextual offer related to that goal. This is a simple trigger, but it must be acted upon immediately.

Similarly, in the financial services industry, a bank may wish to detect if a customer left their money or card in an ATM machine during an interaction – with the goal of alerting them to the mistake as quickly as possible. In a more complex example, a wealth management

firm may combine data from account activity, product utilization, web browsing, and mobile application usage to infer an individual's likely intent at a specific moment. The firm would then use that to present them with more contextually relevant recommendations, messages, and offers.

Data

Accessing and assembling real-time information

A true real-time solution requires the assembly of real-time data, which is used to understand a customer's unique context. Companies interact with individuals across multiple channels and take data from multiple sources. A person's relevant context (the stuff that matters right now) is usually much bigger than what we glean from a single interaction – and it's constantly changing. Failure to look at the bigger picture for each person is where most systems and approaches break down.

For example, you may be able to detect a customer crossing a geofence, which could potentially be important (like when they're entering your retail store, or a competitor's branch location). But that one event is only a small part of a much bigger picture for that customer. While that trigger is important, you need more data to decide whether to take action right now, and if so, what action to take.

For example, if the customer is struggling to use your product, they might be entering the retail store to get help – not to make a new purchase. You wouldn't know that unless you're mining the right data (like call reasons, service likelihoods, and browsing history). Without seeing the bigger picture, you'd probably try to sell them something – at the worst possible moment. That kind of interaction can significantly damage your long-term relationship with that individual, especially if it happens again and again over time.

In many cases, meaningful data is absent or not considered in the decision-making process.

The ability to integrate that additional data is critical to driving an intelligent real-time decision. Unfortunately, data can't sit still in a relational database designed for batch processing; it must often also accommodate data that is "federated" so it can be analyzed and applied in real time.

The five facets of context

You need to assemble data sources that can showcase each aspect of customer context at the specific moment when the interaction is taking place – to enable a truly relevant and compelling interaction.

There are five primary categories of context:

Context	Description	Example Sources
Environmental	Customer's location; where the individual is physically or digitally	Mobile locations, domains, container IDs
Behavioral	Customer's actions; how the individual has interacted or responded	Transactions, interaction history, web browsing
Emotional	Customer's sentiment; how the individual feels currently	Churn scores, default likelihood, call dispositions
Motivational	Customer's intent; what the individual wants to achieve	Propensity to call, call reasons, form submissions
Situational	Customer's state; individual's progress toward achieving that goal	Journey stages, case status, credit risk scores

Many so-called real-time solutions mistakenly skip this aspect of context assembly entirely, jumping directly from an event trigger to a predetermined action.

Consider a simple example: Your cable box stops working so you visit the cable provider's website to find a quick fix. However, this very specific reason for visiting the website triggers an obtrusive popup or marketing offer that stands between you and the outcome that you actually desire (fixing your broken cable box), leading to even more frustration. This breakdown is common and occurs because the brand lacks contextual understanding of that specific customer, even though hints to that context were easily accessible from within the cable box's device data.

Context provides additional perspective with which to make a decision. Examples of high-value contextual data include CRM data, web browsing activity, mobile app clicks, transactions, social posts, device and IoT data, sentiment via text (from chat, social, email, voice-to-text), and third-party data, such as weather or proximity.

In the earlier example of the cable box, the context of the non-functional device serves as an important clue into the situational and motivational context of where this individual might be along their journey – determining what may be the right thing to do (or not do) next.

There are two things to consider when evaluating contextual data:

- **Stagnation:** How quickly does the data stagnate and lose value? People are complex

and constantly changing – and so is their data. Data latency can have a huge impact on relevance, particularly where batch processes are involved. Using stagnant data will lead you to address contextual concerns that no longer exist, such as targeting an on-premises individual with a mobile push offer after they've already left your retail location.

- Scope: Does the contextual data you're assembling cover each facet of context? In many cases, meaningful data is absent or not considered in the decision-making process. Often those insights could have provided a more informed view into the customer's context in that moment and changed the decision entirely. For example, when a brand ignores motivational context, such as a broken cable box, and mistakenly makes a sales offer instead of attempting to fix an active problem.

The speed of data assembly is critical

As soon as a customer need or opportunity is detected (discussed in the next section), the speed of your contextual data assembly suddenly becomes very important. The clock immediately begins ticking and you need to assemble that info with as little latency as possible to make an intelligent, well-informed decision.

Scenario 1: Detection happens quickly, but the data is assembled too slowly – leaving no time to react within the window.

Scenario 2: Data assembly happens quickly, provides enough insight to make intelligent decisions, and still leaves enough time for downstream processes.

Illustration: Latency of data assembly and the window of opportunity

Latency isn't a big concern for certain types of data like mailing addresses, phone numbers, birthdates, or sometimes even product holdings. They simply don't change very often. But mobile app/location data and other digital footprints are extremely fast moving. Was your

customer just clicking on that new gold credit card offer a moment ago? Or were they hovering over the terms for closing their account? Are they still on that same webpage or still in the retail store? Information like this is key to identifying their context.

Context is not only important for hyper-personalization within the channel itself, but also for cross-channel interactions. Without an understanding of their overall context, brands may be caught off guard within a behavioral "blind spot." A customer may not just be traversing across channels, but consuming them concurrently – like when they're viewing terms and conditions on your website while on the phone with a service agent, or viewing product details on the mobile app while shopping in the store. Activities on one channel must inform the next best action on any other channel to optimize that specific interaction and moment. This showcases why "real" real-time capabilities truly matter.

Your decision-making ability is entirely dependent on the data you use to fuel it. In many cases, if your business can't operationalize the contextual data that preceded an interaction, you won't make the right decision for that customer. When that happens, your engagement, satisfaction, and conversion rates suffer accordingly. In today's opt-out culture, every irrelevant message a customer receives moves them a step closer to disengaging entirely.

Decision

How rules can hinder real-time adaptive modeling

Real-time detection and real-time data assembly provide a strong foundation upon which an intelligent decision can be made. However, not all real-time decisions are created equally. Simple "if this, then that" deterministic logic only frustrates customers and gets in the way rather than helping. Those types of rules don't account for the customer's unique needs, emotions, or motivations in that specific moment. Contextual factors like these are the next stage of customer experience evolution for businesses, where a higher level of decision sophistication must be considered.

Levels of real-time decisioning

There are four basic levels of maturity in real-time decision-making:

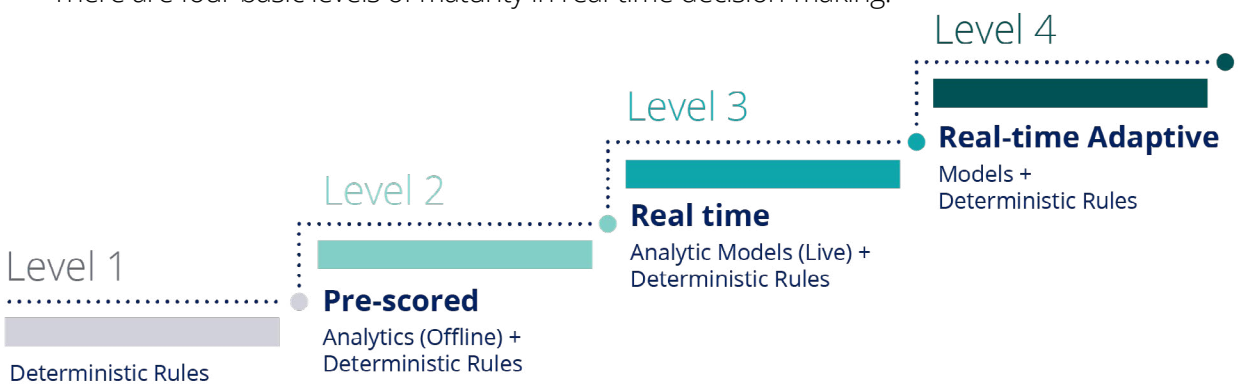


Illustration: Next-best-action decisioning maturity

Level 1: Deterministic rules

Deterministic business rules are often used to configure decision logic. These rules must be relatively simple and straightforward. Otherwise, they become unmanageable and can't determine context and probabilistic outcomes because of all the different factors and permutations to consider. Examples of deterministic rules include if-then-else statements, linear scorecards, and decision trees.

Level 2: Pre-scored analytics (offline) & deterministic rules

This level adds analytic (propensity) models to predict customer behavior. Here, analytic conclusions have been predetermined and pre-scored before the interaction or opportunity even occurs. In these cases, the models are processed in scheduled batches (e.g. nightly/weekly/monthly) based upon then-available data. The resulting scores are then uploaded into the decisioning environment to be used when an interaction occurs. When the time comes, the pre-scored data is used in the arbitration of the real-time decision.

Arguably, this capability is better than having no predictions at all. But there are significant disadvantages that often lead to the wrong decision. Imagine a customer declining a gold credit card offer. This context change should immediately influence the probability to accept other credit cards, lending products, or even the risk of attrition. However, new context will not be taken into account with the use of pre-scored predictions. The scores will not change until the next scheduled update – leading to inaccurate predictions and potentially irrelevant (if not damaging) recommendations.

Consider the prior example of the cross-channel interaction where a customer has just commenced their interaction on the website or IVR with a live agent. Will the analytic models be able to consider those valuable new digital footprints in their calculation and rescore all their conclusions in real time? Pre-scored predictions leave your decisioning blind to this critical context. Offline models are also incredibly inefficient, wasting significant computing power on calculating probabilities that will likely never be used and stagnate rapidly.

Level 3: Real-time analytic models (live) & deterministic rules

A superior architecture will use on-demand, real-time predictions using the freshest context. In this level, the entire decision, from top to bottom, is made in real time to support a next-best-action approach to engagement. The analytic models themselves (rather than just uploaded scores) are executed in "true" real time using the full arsenal of real-time data. Here, the entirety of the next-best-action decision is made using the full context of the customer. The optimal decision for that moment is then selected.

In the cable box example, the Internet of Things (IoT) data from the device – combined with the real-time web click behavior – may be factored into the real-time decision itself, including the predictions that support that conclusion.

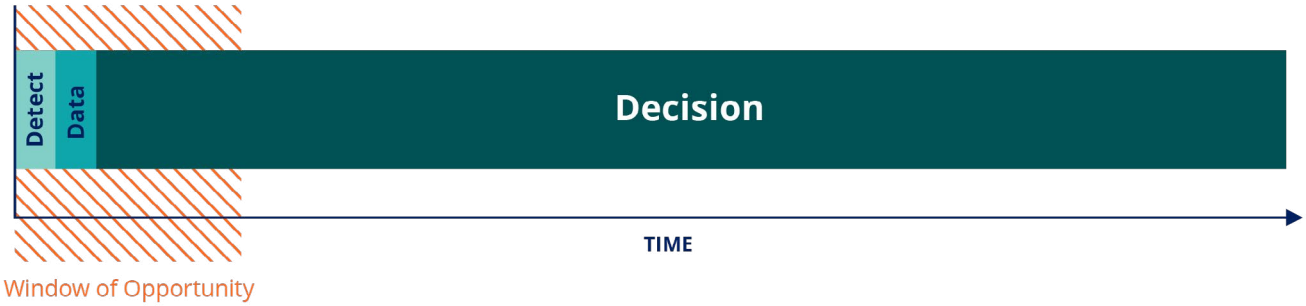
Level 4: Real-time adaptive models & deterministic rules

The most sophisticated examples will combine both deterministic rules with real-time and adaptive (self-learning) models for use in real-time decision-making. Here, the models themselves change continuously based on customer behavior. In this process of machine learning (ML), the models learn from every prediction (correct or incorrect) that they make. When combined with deterministic rules, this results in greater efficiency, more accurate predictions, better decisions, and vastly improved business outcomes and customer experience (which is of course, the whole point).

Speed of decision

Accuracy aside, the speed in which the decision is made is also crucial – as the illustration below depicts. For example, a business detects a potential customer is "in market" for a mortgage and assembles all the data it needs for that decision. But if the business can't make the offer fast enough, the window may have closed and the customer may go to a direct competitor.

Scenario 1: An event is detected and the data is assembled quickly, but the actual decision-making happens too slowly to react within the window.



Scenario 2: An event is detected, the data is assembled quickly, and the actual decision is made in a timely manner, leaving time for downstream activities.

Illustration: Latency of decision process and the window of opportunity

This phase is also where the business' needs and goals are weighed as part of a next-best-action decision, including aspects such as retention, customer lifetime value, service priorities, margin, inventory, opt-out propensity, and much more.

Making decisions at scale

In light of the above, the scale and volume of decisioning must also be considered. For any given interaction, there can be hundreds (or even thousands) of possible actions that must be scored in real time for relevancy, using the full context of that customer. Common decision volumes are 500 – 1000 decisions per second for large enterprises (approximately 86 million decisions each day). Many businesses will require much larger scale than this. Some financial services firms experience peak volumes of more than 200 million decisions daily.

The other key consideration when it comes to decisioning is the scalability of modeling itself. Techniques such as adaptive (machine-learned) modeling allow enterprises to automatically build and maintain more predictive models than they would otherwise be able to achieve with data science resources alone.

Delivery

Responding within the window of opportunity

Delivery refers to the action taken by the brand to respond to the opportunity in-channel, by engaging directly with the customer using the recommended action.

Assessing the point of inflection

It's important to understand that the first three cornerstones – detection, data, and decision – must happen in as real time as possible. This lets the business know whether any action should be taken – and if they are able to do so within the window of opportunity. This serves as an important "point of inflection," as the decision can account for potential latency/irrelevancy in the delivery step. For example, in the case where a business can determine that they won't be able to react within the window with the desired action, or that there is a risk of an irrelevant or negative interaction, the best decision may be to simply do nothing and suppress the potentially belated delivery of an action.

As the time required to make and deliver a decision increases, the quality of the customer engagement and your likelihood of converting an opportunity typically decreases.

For example, a North American financial services firm is detecting events, assembling data, and making decisions in real time at scale, but has also accounted for the option of non-delivery to a customer touch point. When that occurs, the decision is made to record the context of the customer to improve future decisions.

Delivering with speed

Where a response is deemed appropriate, the ability for your business to react in real time is the final mile. The illustration below depicts the impact of latent delivery upon outcome, in the context of a real-time decision.

Illustration: Latency of delivery and the window of opportunity

Speed of delivery can get bottlenecked by the nature of your business, your engagement approach, and the channels you use. You might need to change the way you think about customers and channels in general to maximize your real-time potential. For example, your decision engine might identify a "lead" – a customer with a high propensity to take action right now. You might want your team to make an outbound call to that customer, because they're very good at converting those leads into sales. So you load the lead into an outbound calling queue for a rep to pick up at the next available opportunity: the next morning.

You assembled the data, detected the opportunity, made a real-time decision, and delivered it to the channel – all within a few hundred milliseconds. But you might still be setting yourself up to fail. You chose a channel that was best for your needs, not the customer's. An outbound call can be significantly delayed before that channel actually engages the customer. Not focusing on the customer's timing can translate into missing the opportunity entirely. Many actions may benefit from a faster outreach, such as via real-time SMS, app notification, or email. And that needs to be part of your decision-making process.

Consider this rule of thumb: As the time required to make and deliver a decision increases, the quality of the customer engagement and your likelihood of converting an opportunity typically decreases. Here are some simple examples, at the channel level:

Illustration: Relationship of decision speed and quality of engagement

- On **inbound digital channels** (web, mobile apps, etc.), the need for real-time decisions is extremely high. Customers are constantly consuming content and creating digital footprints that can power models and increase relevance by an order of magnitude, but they stagnate almost instantly. You're forced to fall back to pre-scoring customer needs if you can't make decisions in less than 200 ms, which would cause the likelihood of converting the customer to drop astronomically.
- On **inbound agent channels**, like the call center or in a retail location, the need for real-time speed is still very high, but the performance service-level agreements (SLA) drop slightly. Live conversations have a slower cadence, with natural pauses built in. In these types of channels, you typically must return decisions between 500 ms to one second in order for the agent to process and use them in the natural course of conversation. Any longer than that and those pauses become awkward, the agent stops using your recommendations, and they go back to relying strictly on their own subject matter expertise – which ignores most aspects of customer context altogether.
- In **digital outbound channels** like email, SMS, or mobile push, speed is important, but it's no longer measured in milliseconds. It's typically measured in seconds or minutes. For

example, geolocation data around an airport, train station, or retail location may identify a great window of opportunity to engage a customer and uncover a significant need. However, you don't have to make that decision in under 200 ms. Typically, that window lasts anywhere from a few seconds to a few minutes. Any longer than that, though, and the customer moves on. When the window closes, your outreach becomes far less effective and might even become creepy or intrusive.

Understanding "rift vs. reward"

When assessing your real-time capabilities, there may be a big difference between "rift" (the time span keeping your business from being in the game) and "reward" (the opportunity to actually win that game). Your rift or reward is cumulative. Each delay across the four categories compounds and impacts additional downstream steps. If you overshoot the opportunity window, you lose out on the opportunity itself.

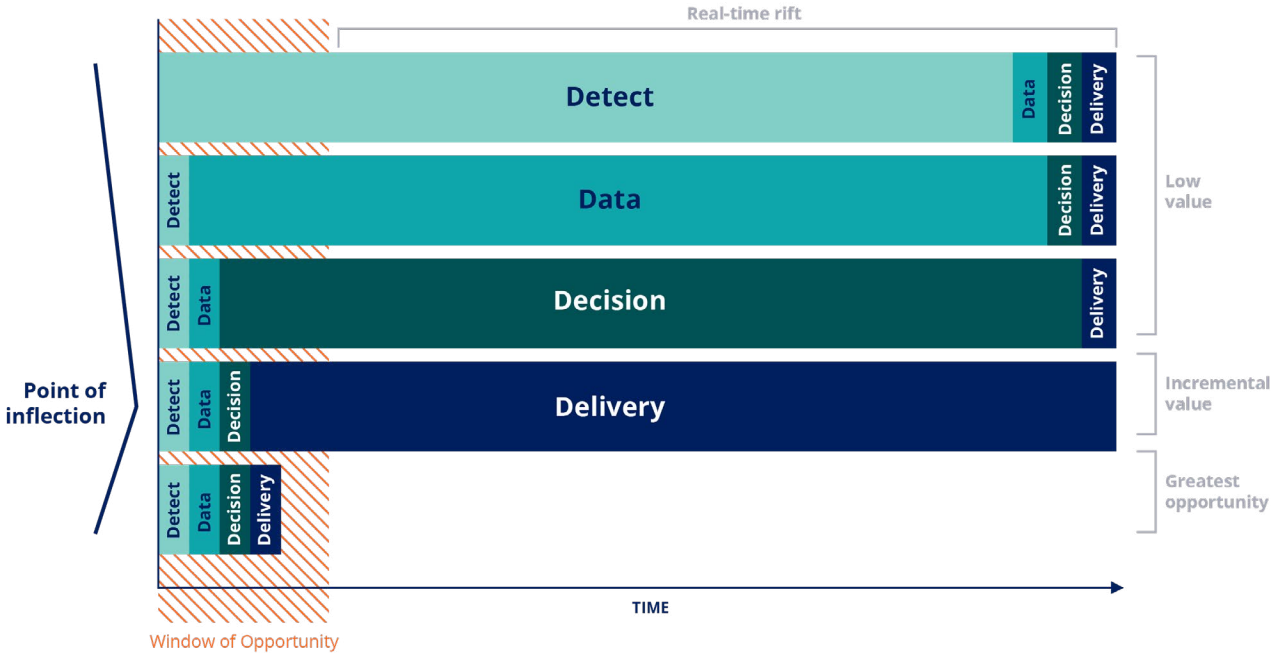


Illustration: Understanding the impact of real-time bandwidths

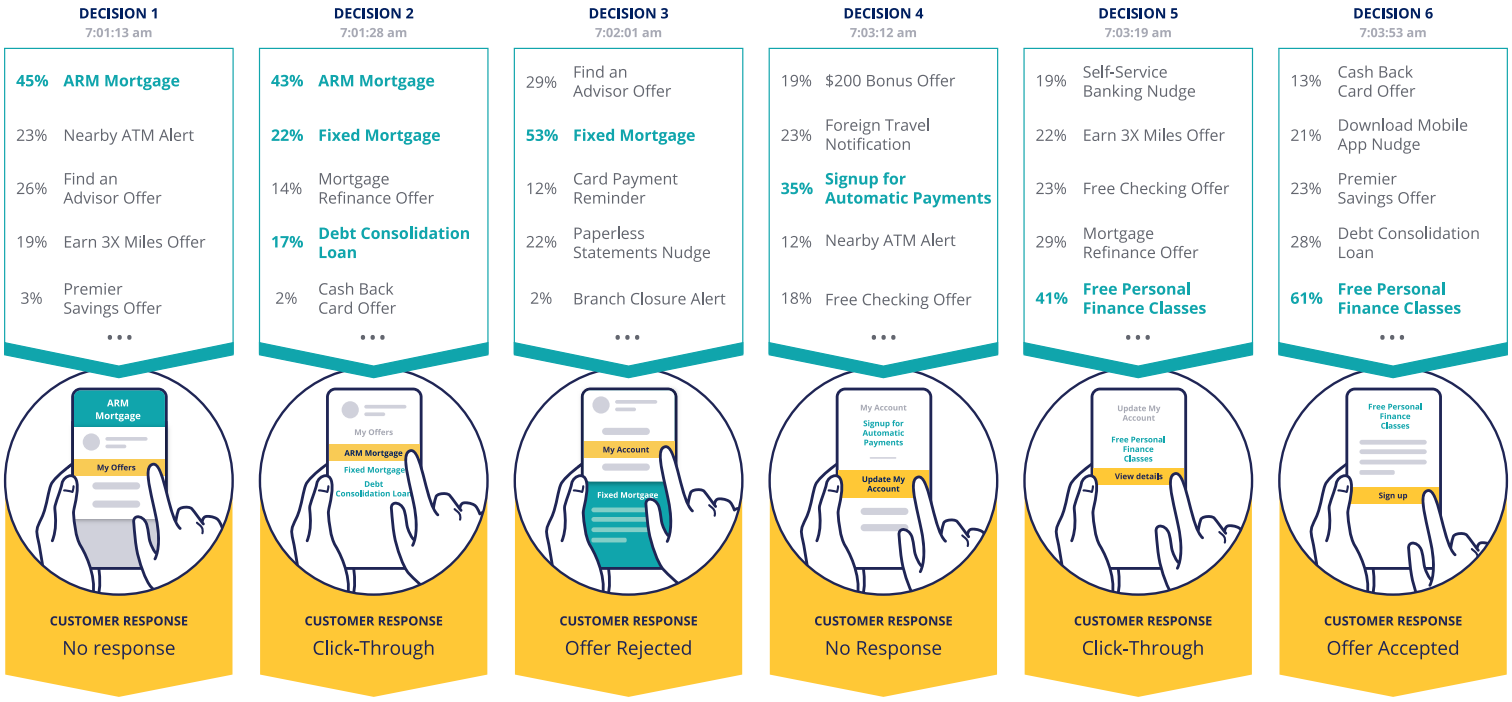
How fast is fast enough? If you can't adapt right along with the customer and their decision-making process, you're out of luck. You're wasting the opportunity to use that moment for value creation and putting the customer experience at risk.

Aggregators and new intermediaries are constantly inserting themselves between you and your (potential) customer. You're not just competing against the clock; you're competing with each of your rivals, too. It's not just a matter of who will respond, but who will respond first with the most compelling message. Within this window of opportunity, the space between your responsiveness and the next-fastest competitor is precisely your opportunity for competitive advantage, and potential for "real-time reward."

The “real-time rift” is the gap between optimizing an opportunity, and not having one at all. The most sophisticated next-best-action solutions will “re-decision” the entire real-time process multiple times during a single, live customer interaction. This can mean making anywhere from 10 to 50 distinct decisions within just a few minutes, all within a single customer experience. Each time, the propensity scores are re-calculated for each potential action, taking into account the new data and context. One of the USA’s largest telecommunications carriers applies this approach to real-time customer retention scenarios; if a customer calls to cancel, Q&A is recommended to better understand context through data. The responses drive the retention offer and each additional calculation supporting it. In these cases, new data shapes the next step throughout the interaction, processing new data all along the way.

Re-decisioning example

The figure below shows the path taken by an individual customer as they progress through a typical mobile interaction with a brand leveraging real-time decisioning. Over the course of just 2 minutes and 40 seconds, the brand “re-decisions” that customer profile six distinct times, each time accounting for the new mobile click-through data and other behaviors demonstrated by the customer during the experience. The end result is that the individual is presented with six very different but highly relevant next best actions, as they navigate through the mobile app, and shift their context.



To truly understand the power and complexity of the decision-making happening in this example, it's important to know the following:

- A brand that can effectively "re-decision" its customers must maintain a **library of potential actions** to draw from – including a mix of sales offers, retention recommendations, proactive service "nudges," and nurture messages.
- Enterprises often have **1000+ actions** that may be in play for a single customer, at any decision point. Each of those actions also has multiple creative treatments associated with it, representing various combinations of imagery, text, color, channel application, etc.
- Each action and treatment is also associated with a series of **propensity models** – these models are re-scored with new data every time a decision is made. Those models' scores are used to arbitrate across actions to determine what is most appropriate for the customer, in that exact moment, given their full customer profile and their current (seconds-ago) context.
- Each click-through, hover, offer rejection, etc., represents a **change to the customer's context.** When those events occur (like when a customer selects "no thank you," or simply navigates to a new page), that data is immediately activated for real-time use, mapped to the data model maintained in the application, and that profile is re-scored to calculate a new "next best action."
- So in the above example, as a customer is presented with actions and selects/rejects/ignores them, and as their click-through behavior is aggregated and curated, the brand is **constantly shifting their approach.**
- The decision engine adapts from recommending sales offers in the first three interactions, to instead making service recommendations, retention, and nurture-type actions for the rest of the experience, because the individual's behavior changed the propensity scores, and showed those "softer" actions to be more **relevant to their context and intent.**
- While brands have traditionally attempted to align with customer context using only segmentation or business rules, those are legacy "one-to-many " approaches that **fail to adapt for real-time context,** or achieve the desired level of relevance. So they typically fall flat.
- Without a propensity-driven approach that's constantly re-decisioning each customer profile based on new data, it's impossible to achieve the **5X increase in relevance** that's associated with a real-time investment.

The most sophisticated next-best-action solutions will "re-decision" the entire real-time process multiple times during a single, live customer interaction.

Conclusion

Not all real time is created equal. Selecting a next-best-action solution that cannot support all four steps in "real, real-time" puts your customer experience and revenue potential at risk. Independent analysts within industries, such as telecommunications, financial services, and insurance, have found that on average, a successful implementation of real-time next best action delivers more than $225 million USD in incremental sales and retention value for every 10 million customers, per year.

For every month your business is not using a truly best-in-class next-best-action solution that is capable of real, real-time capabilities, you are likely costing your business in the order of $20 million for every 10 million customers that you have.

It's important to start simple in your strategy – but don't make the mistake of starting simple with your software. Your real-time, next-best-action foundation needs to be able to handle all of this complexity. Choose wisely and you'll soar ahead of your competion. If you don't, you'll quickly find yourself out of runway.

It's time to get serious about your real-time future.

Pega is the leader in cloud software for customer engagement and operational excellence. The world's most recognized and successful brands rely on Pega's AI-powered software to optimize every customer interaction on any channel while ensuring their brand promises are kept. Pega's low-code application development platform allows enterprises to quickly build and evolve apps to meet their customer and employee needs and drive digital transformation on a global scale. For more than 35 years, Pega has enabled higher customer satisfaction, lower costs, and increased customer lifetime value.

For more information, please visit us at **www.pega.com**