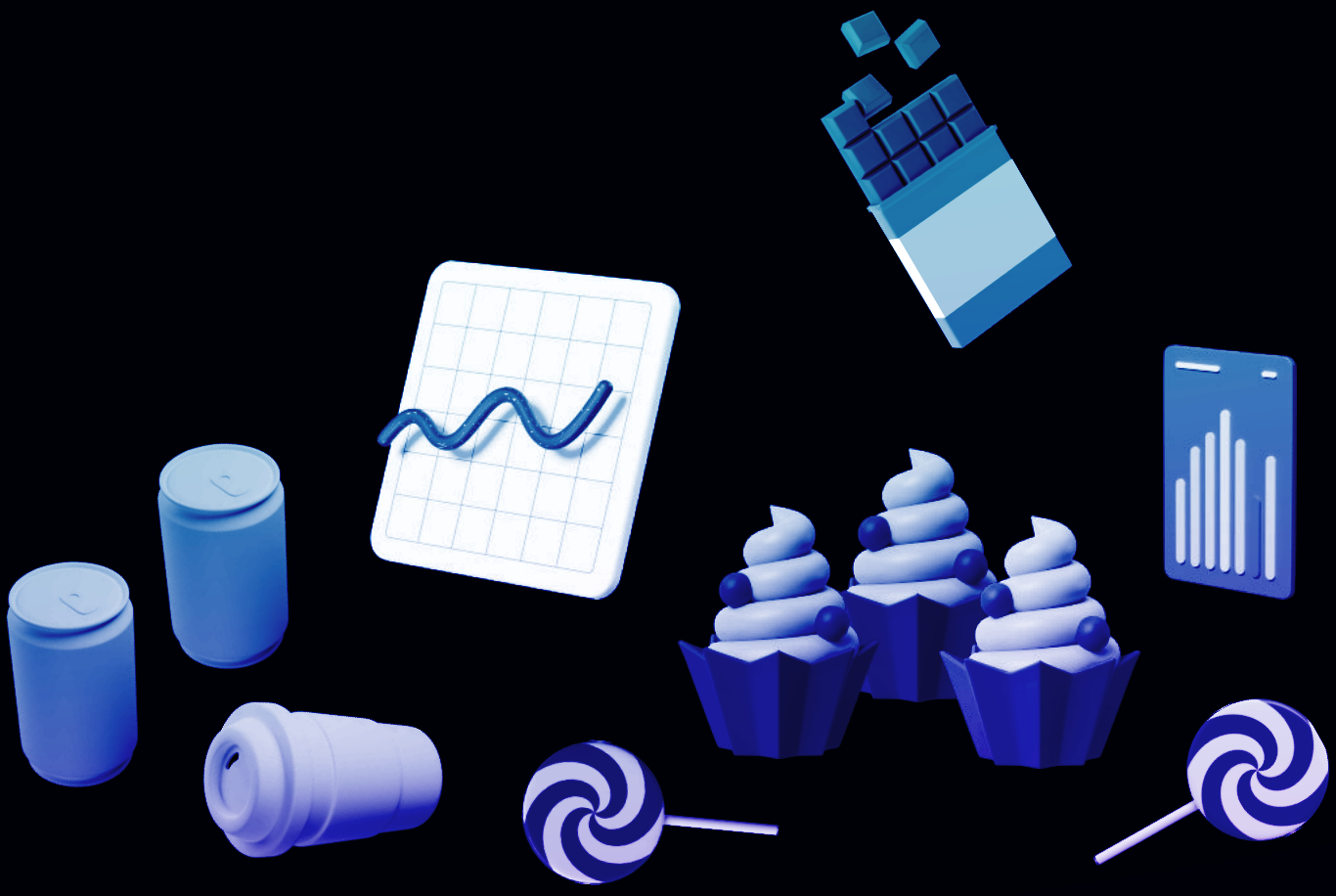
DUALITY

Is Your Machine Learning Model Bingeing on Junk Data?

The Three I's Framework for Realistic and Relevant Synthetic Data



Contents

Is Your Machine Learning Model Bingeing on Junk Data?

An Introduction to the *Three I's Framework* for Quantifying the Realism and Relevance of Digital Twin Generated Synthetic Data.

Wouldn't it be nice to be able to just push a button and have all the data you want at your fingertips? In theory, that is the promise of synthetic data: to solve one of the biggest challenges of using Machine Learning (ML), the collecting and labeling of relevant data, simply by generating data in a simulated environment and then using it to train your ML models. But it is not easy to capture all the complexities of the real world, so synthetic data and simulation frequently fall short of their promise, leading to poor ML performance once released into the wild of the real-world data. The accumulation of all the differences between real-world gathered data and its synthetic counterpart is known as the **Domain Gap**, which can be large and multi-faceted – making it very difficult to identify and address those discrepancies which contribute to poor performance. Another way to look at this problem is that training a model on poor quality data, i.e., one with a large domain gap and a high percentage of irrelevant information, can cause the model to pick up "bad habits" that don't transfer to real data.

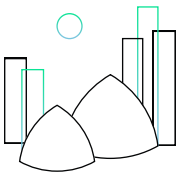
If we are to embrace the promise of synthetic data, addressing the Domain Gap issue is a crucial step. While there is a range of currently favored approaches for closing the Domain Gap, at Duality we leverage high quality Digital Twins, and believe this represents a thoughtful, systematic and future proof approach to generating high quality synthetic data that in turn results in an impactful and predictable return on data investment.

What are Digital Twins?

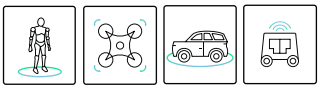
> A digital twin is a virtual representation of real-world entities and processes, synchronized at a specified frequency and fidelity.

Source: Digital Twin Consortium

Simply put, Digital Twins are highly realistic digital versions of real-world entities. The primary purpose of a Digital Twin is to accurately present the appearance, properties and behaviors of a physical object in a virtual setting. To achieve this goal, Digital Twin acquisition requires meticulous 3D modeling, high quality real-world gathered data that sufficiently describes essential aspects of the entity, and state-of-the-art physics engines to integrate it all together. To that effect, a Digital Twin can be generated from a boundless pool of real-world sources with ever increasing complexity: a single flower stem and a field of wildflowers can both be represented as different types of Digital Twins. We classify Digital Twins into three basic types: **environments**, **systems**, and **items**.

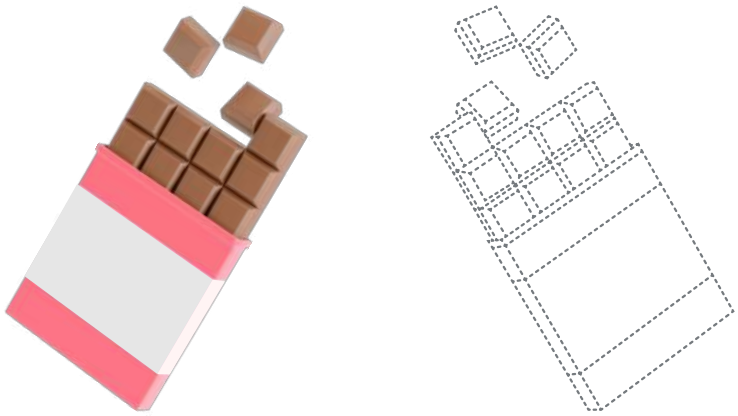


Environments are the encompassing surroundings in our domain of interest – they can be as broad as a forest or the streets of a city, or as narrow as a particular spot on a conveyor belt.



Systems are any entities that perform or exhibit behaviors in the environment.

Items are any non-functional objects or products that populate the environment and that systems can interact with.



Digital Twins for More Precise and Efficient Training

We have observed that highly realistic and relevant synthetic data can match and augment real world data leading to robust and deployable ML models. This implication, that the quality of our synthetic data helps predict successful ML model training, leads us to pose a question:

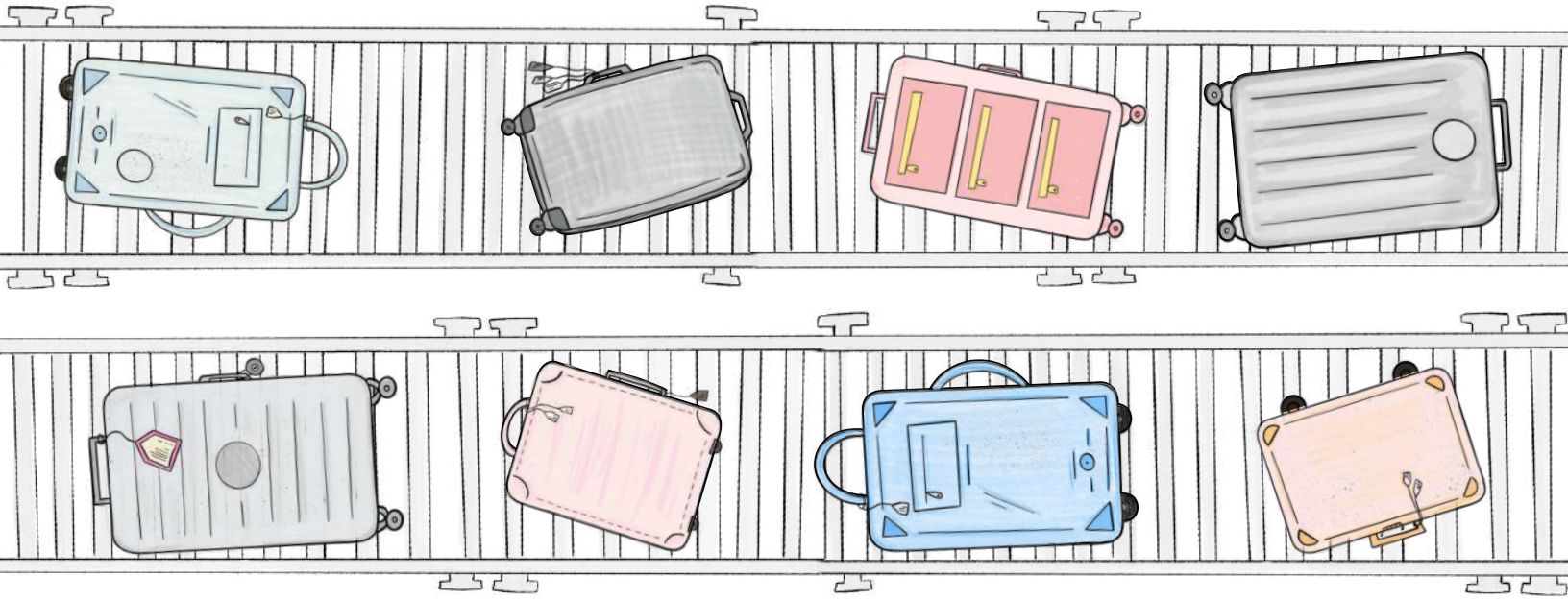
How do we quantify the realism and relevance of our Digital Twins, even before using that data to train a model?

To this end, we came up with three criteria to guide the creation of synthetic data. They are collectively referred to as “The Three I’s”: **Indistinguishability**, **Information Richness**, and **Intentionality**.

Introducing the Three I's

1. Indistinguishability

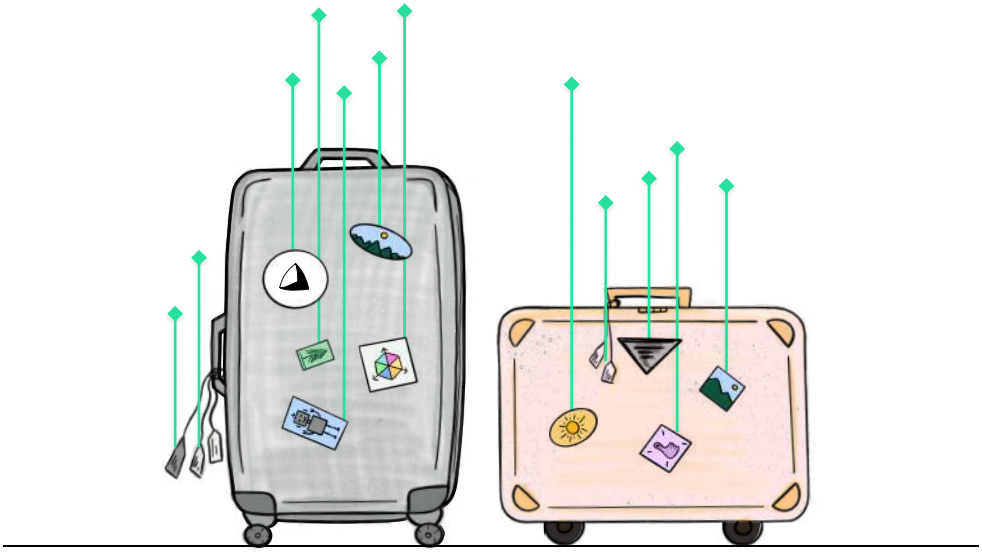
The first step towards good synthetic data is minimizing the Domain Gap. Therefore, our synthetic data should strive to be indistinguishable from a real-world sample. It is not supposed to be identical, but it should be impossible to determine if any given distribution of data came from our simulated version or from a real-world example. In other words, an impartial algorithm sorting data as either 'real' or 'fake' should be wrong at least 50% of the time – the real-world samples should completely blend in with the synthetic ones. **The higher our Indistinguishability rating, the more precisely our data will capture a specific scenario.** We will expand on how we evaluate the Indistinguishability of Digital Twins in the next section.



Indistinguishability: In this example, the suitcases aren't identical (synthetic ones are mixed in with the real ones) and we cannot tell which is which.

2. Information Richness

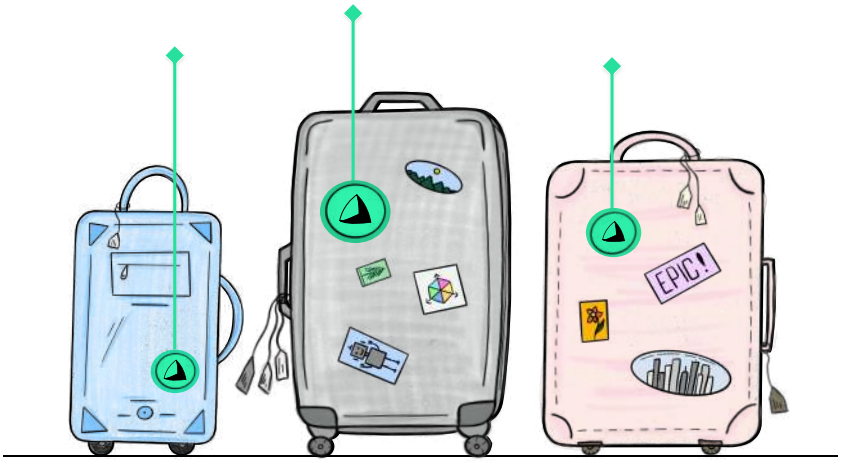
While synthetic data should be indistinguishable according to the metric outlined above, it also needs to be novel – to be useful, it needs to be generating new information about a specific domain. The data should provide, for example, new perspectives, new angles, new features, etc., that fill in the gaps of the real data. We don't want to dilute the data set with redundant information, so each data point should be valuable and representative of the real-world scenario. If Indistinguishability allows for high precision, **Information Richness allows us to accurately broaden the horizons of what our data can capture.**



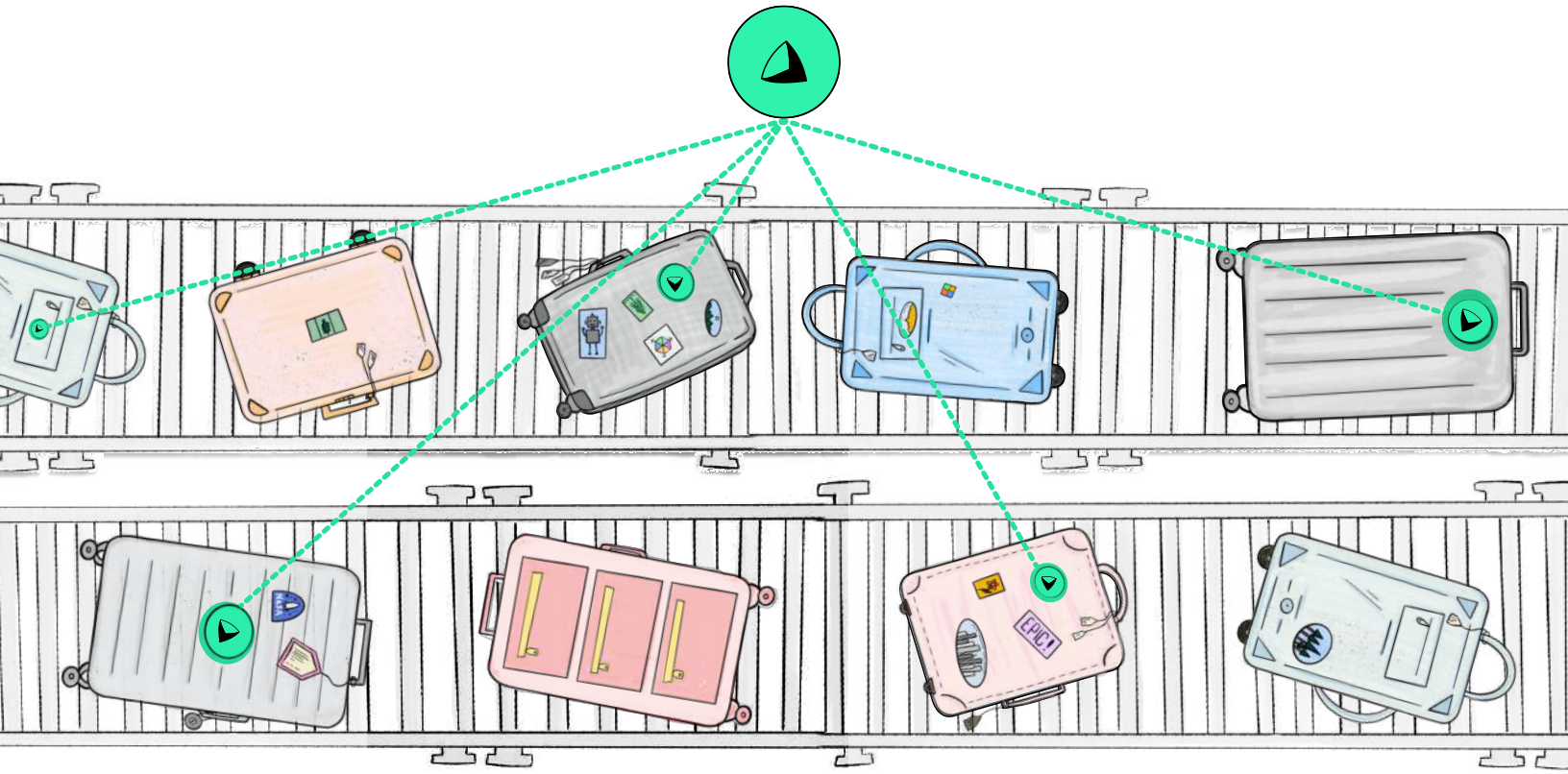
Information Richness: Generating new versions that we did not see in the wild. Here, we have new suitcases covered in infinite varieties of stickers and tags in familiar and novel variations.

3. Intentionality

We are seeking to have a fundamental understanding of the data we are simulating and what aspects of it are useful in our domain. In generating new data, we want to identify key items so that we can create variety in the most relevant variables. **Through Intentional data, we define our domain of operation.** In other words – even though we can create infinite amounts of variation in our synthetic data, not all variations are useful for improving the performance of an ML model. Simply introducing Information Richness without consideration for the use case, or relevance to the model, often yields results that are either negligible, or potentially confounding. Thus, to create a robust ML model, we can make a clear decision on its intended domain of operation: which conditions are relevant? Are they variable or static? What edge cases are significant, and which ones can be ignored? Intentionality strives for the holistic understanding of what specific Information Richness to introduce and can be viewed as the control lever for how far and in what directions we venture from our homebase of real-world gathered data, for any given use case.



Intentionality: Identifying our domain of interest. For this hypothetical model we focus on suitcases that feature the Duality logo sticker.



“The Three I’s” are intertwined and interdependent. To help us visualize these abstract relationships we may imagine that any domain can be represented as a unique three-dimensional shape. Indistinguishability is the structural core of this shape, where our real and synthetic samples blend together. Information Richness is how much of the domain we fill, or all of the ways the shape can evolve away from the core. Intentionality is then the guide of this evolution, pruning the irrelevant and highlighting the valuable aspects, ultimately defining what the shape looks like. As we come to better understand our domain of interest, we are better able to dictate exactly what Information Richness is introduced, and the shape takes on a clear and defined form.

As an analogy, we can imagine building a house from a set of general blueprints. Much of the original structure would closely line up with what the blueprints dictate (Indistinguishability) – it is how we know that our house is safe and functional. Information Richness then, is all the ways we can evolve the construction: the materials we can choose, the rooms we can partition or add on, the appliances we can install, etc., - all things not expressly dictated by the blueprints, but guided by their foundational layout. Intentionality is how we determine which of those choices are right for our specific house: what features are important for our location? Or for the climate? Or for the number and type of occupants? All these choices evolve the house away from the original blueprints, but in ways intentionally consistent with their boundaries and our needs. To summarize this idea in the world of synthetic data: we build on top of our initial structure by making informed choices that intentionally increase Information Richness to guide the direction of our ML model.

Considering all of the above, we think it is fair to say that providing increasingly Indistinguishable, Information Rich and Intentional synthetic data is possible and necessary.

Evaluating Indistinguishability

As we mentioned at the outset: high quality synthetic data can match and even outperform real world data in ML model training. This correlates high Indistinguishability with better training results, and mandates that we quantitatively evaluate Indistinguishability before we subject any model to a particular data diet.

The first step of evaluating a Digital Twin scenario is to evaluate the Digital Twin items individually. For example, if we are training an object detection model, we will first evaluate the indistinguishability of each item/system individually, followed by repeating the process in the intended environment. We will walk through the evaluation of an individual Item Digital Twin, but the same process is followed for digital twin systems and environments as well as Digital Twin system/items in their environment.

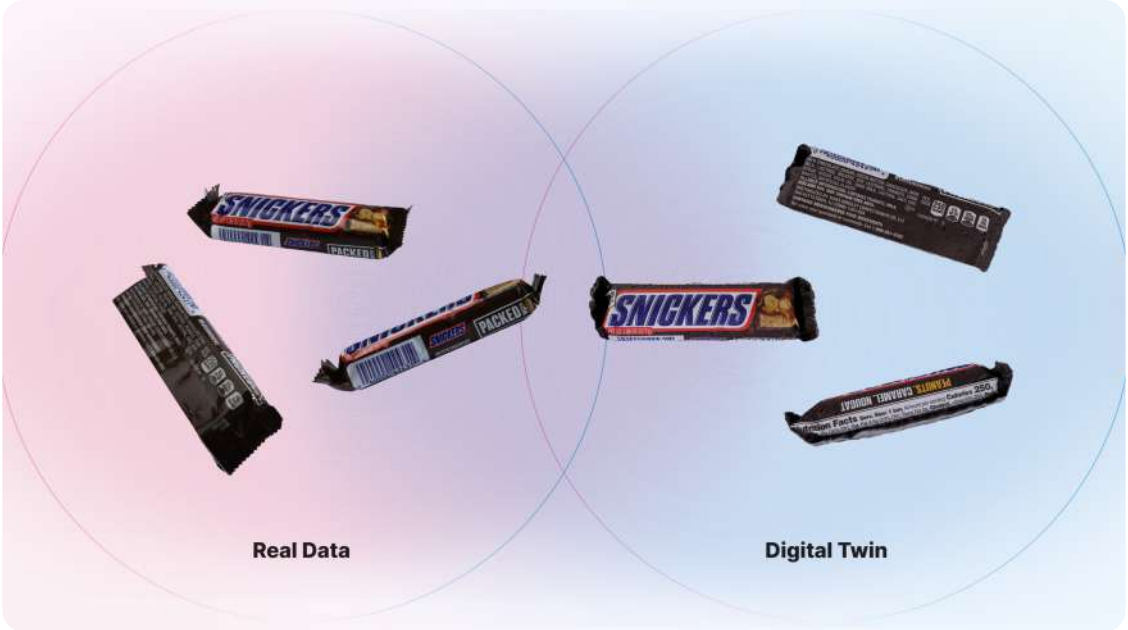


Fig. 4 Example of Real Images and Digital Twins.

As shown in Fig. 4, we use images from the Digital Twin we made and the real-world object to evaluate their Indistinguishability. Our friends at Voxel51 have developed an open-source software tool called FiftyOne that supports visualization and analysis of data sets in machine learning.

We leverage FiftyOne to facilitate the calculation and visualization of the Indistinguishability Score (we present an example below, but encourage anyone to try FiftyOne and our repo on their own data). In order to represent the visual aspects of the image as quantitative features that can be analyzed, the images are sent through a pre-trained convolutional neural network. From here, FiftyOne's implementation of dimensionality reduction is used to visualize the data.

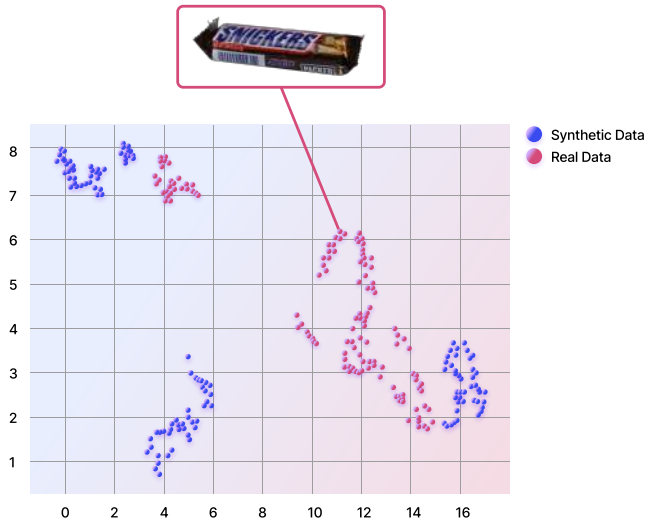


Fig. 5a Distinguishable synthetic data

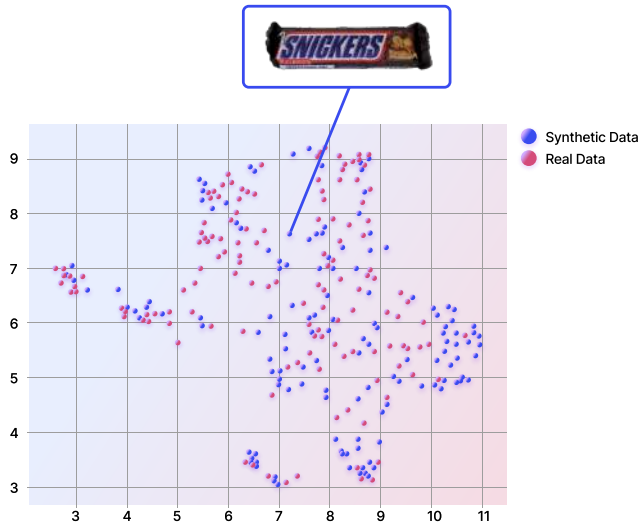


Fig. 5b Indistinguishable synthetic data

In Fig. 5, each data point represents a unique real (blue) or synthetic (red) image. Here, we are presented with equal amounts of real and synthetic data. If the data are indeed Indistinguishable, then the likelihood that the closest sample next to a random synthetic image is real should be 50%. In other words, the synthetic data images are perfectly mixed in with their real-world counterparts, and the distribution of synthetic data Indistinguishably matches the distribution of real-world data. Please note that ideally the real world data should not change between Fig. 5 a), b) but since dimensionality reduction is an iterative process it needs to be done on synthetic and real data at the same time. This is why, although the real data in 5 a), b) is the same, the different synthetic data causes a different low-dimensional representation [T-SNE] [UMAP]. This technique is not limited to any specific dimensionality reduction and can benefit from other types of dimensionality reduction.

In the cases when the data are not completely Indistinguishable, we can follow the same logic by generating a 'data overlap value'. This value quantifies how much of the synthetic data is actually overlapping with the real data and represents it as a succinct Indistinguishability Score. In this scenario the probability of a synthetic sample having its nearest neighbor be a real-world sample will be less than 50%. Once we have this probability, we can calculate how much this distribution deviates from fully Indistinguishable to get the Indistinguishability Score.

This same approach can also be extended to cases when there is more synthetic than real data — which just happens to be all the time! If we have 70% synthetic data and 30% real data, and the data are Indistinguishable, then the likelihood that the closest sample located next to a random synthetic image is real should be 30%. Again, this probability is then converted to a Indistinguishability Score that now accounts for large imbalances in amounts of real and synthetic data and for data that are less-than-Indistinguishable to give us a good estimate of the realism of our synthetic data.

In Fig. 5a and 5b, there are an equal number of synthetic and real images, which means that synthetic data should be nearest to a real data point 50% of the time. In Fig. 5a, this is observed 0% of the time, while it is observed 40% of the time in Fig. 5b. With these observed probabilities, we estimate that the data overlap is 0% and 80% for Fig. 5a and 5b, respectively. Check out our repo to test these methods on sample data or try it out on your own data.

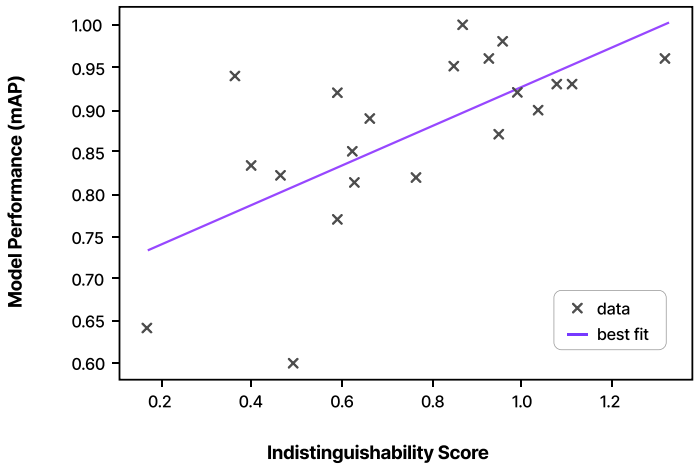


Fig. 6 Indistinguishability score (data overlap) vs mean Average Precision (mAP) for object detection. Note how a higher Indistinguishability Score predicts better model performance.

Fig. 6 demonstrates the relationship between the Indistinguishability Score of a Digital Twin and the performance of models trained on that Digital Twin. In our testing, we have observed that there is a clear positive correlation between Indistinguishability Score and the performance of object detection models. What's more, we can also see that a Digital Twin does not need to be perfectly Indistinguishable in order to yield significant benefits to an ML model. In fact, a score greater than 0.8 does not necessarily produce an improvement. The reason for this is that Indistinguishability is not the only important factor - many factors impact model performance using synthetic data, and top among these is a direct tension of Indistinguishability with Information Richness.

While the specifics of Digital Twin acquisition is beyond the scope of this paper, it is important to remember that Indistinguishability is always rooted in the measuring and analyzing of a real-world sample, and is an approximation of the relationship with the actual real world. Furthermore, even if we have access to complete real-world data, it does not mean that it is the distribution that we want. For example, we may want to oversample edge cases that are not common in the real world but are very important for good training, therefore real world data may not be the end goal. This is why all Three I's are essential for optimal synthetic data.

In the following section we address these points, and dive deeper into Information Richness and its tension with Indistinguishability. We break down how we conceptualize Information Richness, postulate its usefulness, and explore how it fits in with today's dominant methods of creating data diets for any ML model.

Evolving Information Richness

As noted previously, Indistinguishability is only one part of the methodology that leads to good training outcomes. In this section we explore the second key component: **Information Richness**.

The reason why the Indistinguishability Score alone is not sufficient to evaluate the quality of our data is that we are not only trying to replicate real-world data, but generate new information beyond what is gained from a few real-world samples. If our data is Indistinguishable, but provides no new information for the model, then the data is simply redundant – and what would be the point of generating it in the first place? A model needs to encounter a vast multitude of novel and diverse iterations of the data to prepare it for the unpredictable conditions it will encounter in the real world. Thus, variety in large scale data sets is the platonic ideal of good synthetic data. At Duality, we believe that our approach to this synthetic creation of variety (i.e., Information Richness) adds true value to our more Indistinguishable Digital Twins.

What is Information Richness?

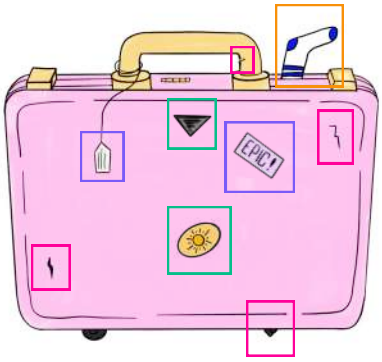
Simply put, Information Richness is a measure of novelty or uniqueness within a synthetic data set. In general, data does not need to be useful or realistic to be Information Rich; it just needs to be highly varied. For our purposes, we conceptualize it as the expansion of the data distribution from a real-world sample to fill-in and expand a simulated domain. If appropriately applied, it allows us to accurately broaden the horizons of what our data can capture and better simulate our domain of interest.

Recalling the Snickers candy bar example from the previous post: we generated images that were as Indistinguishable as possible from the real-world gathered images. This could have also been done by duplicating the real dataset, but this would not increase the Information Richness. To make a set more Information Rich, we must generate additional images that vary from the original real-world samples.

To name a few options, we can create images from different angles, change the lighting conditions, or even deform the Digital Twin itself. When we generate images that are visually close to our real-world sample we are filling in the gaps of our domain, as these are images that we could have likely observed by collecting more real-world data (Ex. 1). This is not unlike how a cartoon animator fills in all the in-between poses of a character between two keyframes. Alternatively, when we generate images farther afield from our real-world sample, ones that present conditions not commonly observed (Ex. 2), we are expanding our domain and creating all sorts of edge cases (realistic as well as extremely unrealistic). Both of these scenarios increase Information Richness, but the latter also decreases the Indistinguishability Score of the images, and we expand on this in the next section.

Two schemes for generating Information Richness:

Ex. 1: Filling in the gaps of the real-world sample.



Ex. 2: Highlighting and oversampling edge cases.

Not All Information Richness Is Created Equal

The current dominant approach for creating large, information-rich synthetic data sets for ML training is called **Domain Randomization** (DR). With DR, a set of parameters is randomly altered to create a much larger set of novel synthetic data. Randomness is the operative concept here as DR theoretically produces such a wide swath of cases that it blanketly fills the blind spots we might have in our Domain Gap. In practice, this means that the immense data sets produced contain everything from data that is similar to the domain of interest to data so far removed from reality that the ML model would never actually encounter it. In between those two extremes, the model is theoretically exposed to a sufficient amount of relevant data that generally captures most real-world scenarios.

However, since these data are randomly generated, without regard for Indistinguishability or Intentionality, their Indistinguishability Scores tend to be quite low. Images of Snickers candy bars placed on the moon with astronauts (a hypothetical example of potential DR generated data) certainly increase the Information Richness of a data set, but they also decrease the Indistinguishability Score as they don't reflect the reality that our model will encounter in the real world. This crossing of the boundary between edge cases and impossibility contributes to inefficient training and is a simplified example of the natural tension that exists between Indistinguishability and Information Richness (Fig.1).

Since Indistinguishability is always rooted in the measuring and analyzing of a real-world sample, the Indistinguishability Score of novel synthetic data is always tied to that sample and will decrease as soon as new data starts to drift away from the core distribution of the aforementioned sample. In other words, Information Richness that expands the domain is also more speculative. But this expansion of the domain is essential for capturing all the varieties of cases relevant to successful ML model training.

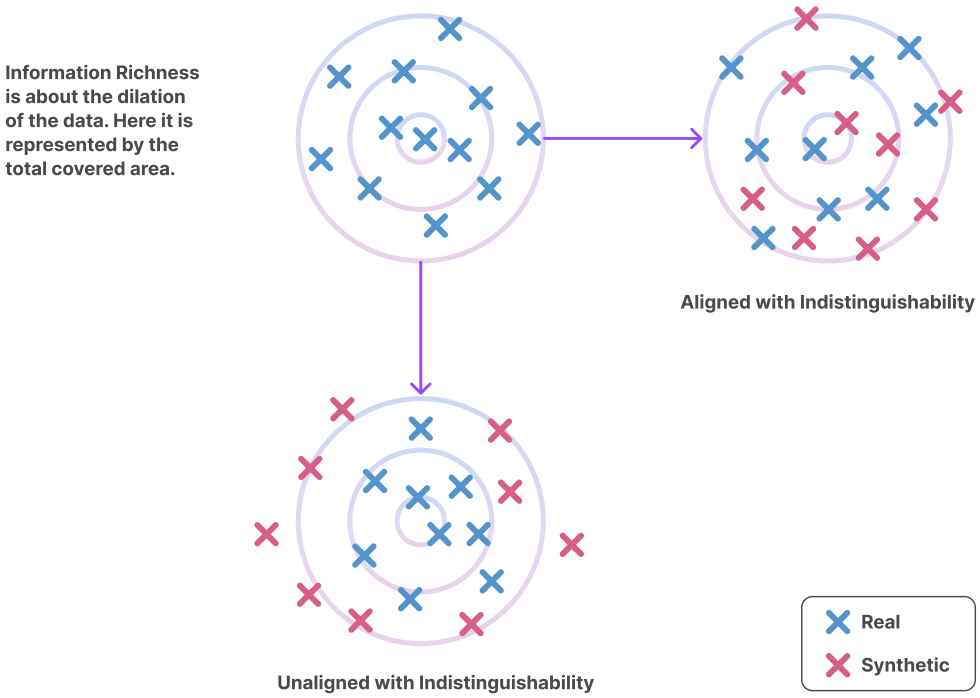


Fig. 1 The natural tradeoff of between Information Richness and Indistinguishability

Does this mean that increasing Information Richness away from the core distribution creates less realistic synthetic data?

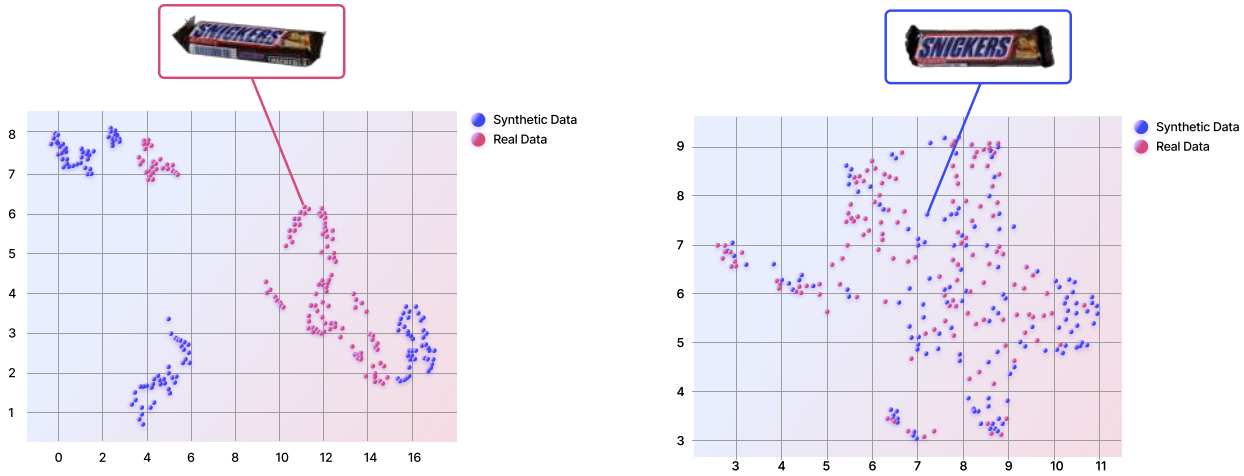
Not necessarily.

This tradeoff between Indistinguishability and Information Richness is not universal nor zero-sum. Real-world data, if it were possible to be gathered in mass, would always be highly Indistinguishable and highly Information Rich. Thus, we can postulate that many synthetically created samples should have real-world counterparts; they just weren't the ones that we observed. Moreover, edge cases are by definition not common in the real world, but are crucial for successful training. An ideal training dataset over-represents these pivotal edge cases while in the process sacrificing some Indistinguishability. This is where it is important to highlight that for training efficacy, emphasis on Indistinguishability has limits. So while we will always have to balance Indistinguishability and Information Richness in useful synthetic data, they are by no means operatively opposite of one another.

Evaluating Information Richness

When we evaluate Indistinguishability, we use data clusters in which individual points represent 'real' and 'synthetic' images. The alignment of 'synthetic' to 'real' clusters is a key indicator of Indistinguishability: the higher the overlap between the clusters, the more Indistinguishable the synthetic data is from the real-world data. With Information Richness, we look at the area these clusters cover. The more expansive the 'synthetic' data cluster is, the more Information Rich the data set is. More simply put: **Information Richness can be measured by the area of the cluster.**

To illustrate, we can revisit the Snickers bar example from the previous section. Just as with calculating the Indistinguishability Score, we start by applying a Convolutional Neural Network followed by dimensionality reduction. Here, we are again leveraging FiftyOne, an open-source software tool developed by our friends at Voxel51. This provides us with a 2-d representation of each image, as shown in Fig. 2.

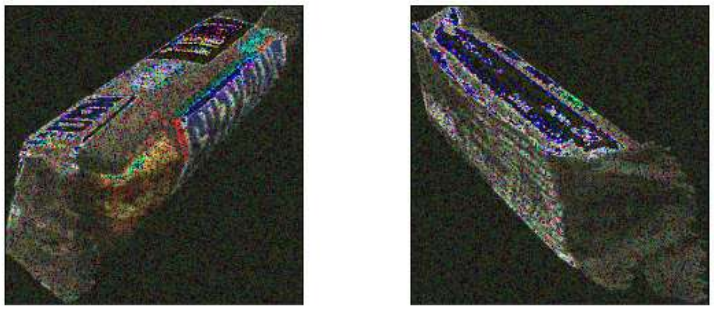


Distinguishable synthetic data

Indistinguishable synthetic data

Fig. 2 Not only does the dataset on the right have higher Indistinguishability but it also has higher Information Richness.

But as we stated above, for Information Richness evaluation, we are not interested in points but in areas. To calculate the area of the cluster, we must first assign an area to each point. To do this, we apply random noise to approximate variability in the images (Ex. 3). This is done once per image and the difference between the randomized image and the original image is then used to create a scatter plot as shown in Fig. 3. This scatter plot is made up of all the images and each point represents an image pair consisting of the original and randomized image. We then use this scatter plot to calculate an area radius by finding the median distance from zero of the points; in Fig. 3, the median distance from zero is 0.51.



Ex.3 Images of Snickers bar Digital Twins with random noise applied to approximate variability.

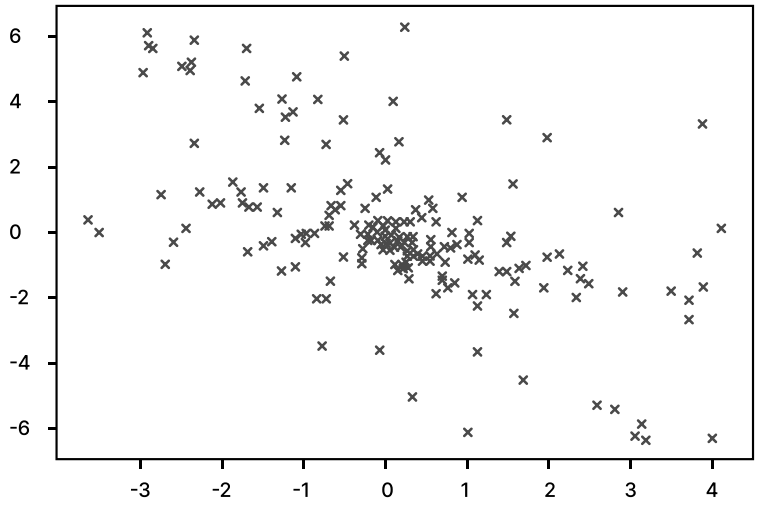


Fig. 3 Scatter plot of randomized real and synthetic images minus nonrandomized images. Center represents the location of nonrandomized images.

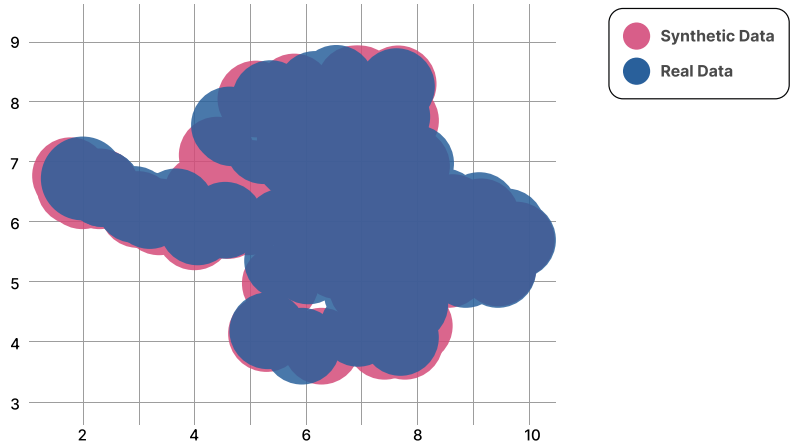


Fig. 4 Information Richness of synthetic and real data.

Once we have this radius, we can expand each of our original points into circles that approximate the area coverage of each image. Fig. 4 visually demonstrates how these points from Fig. 2 are now converted into areas. The aggregation of all the areas of individual circles creates a total area for the point cluster and we can then calculate the overlap between real and synthetic data as well as how much unique area synthetic data is providing. In Fig. 4, we can see that the synthetic data is providing greater Information Richness than the real data sample while still exhibiting a similar distribution, meaning that we have gained new information while still keeping a high Indistinguishability Score.

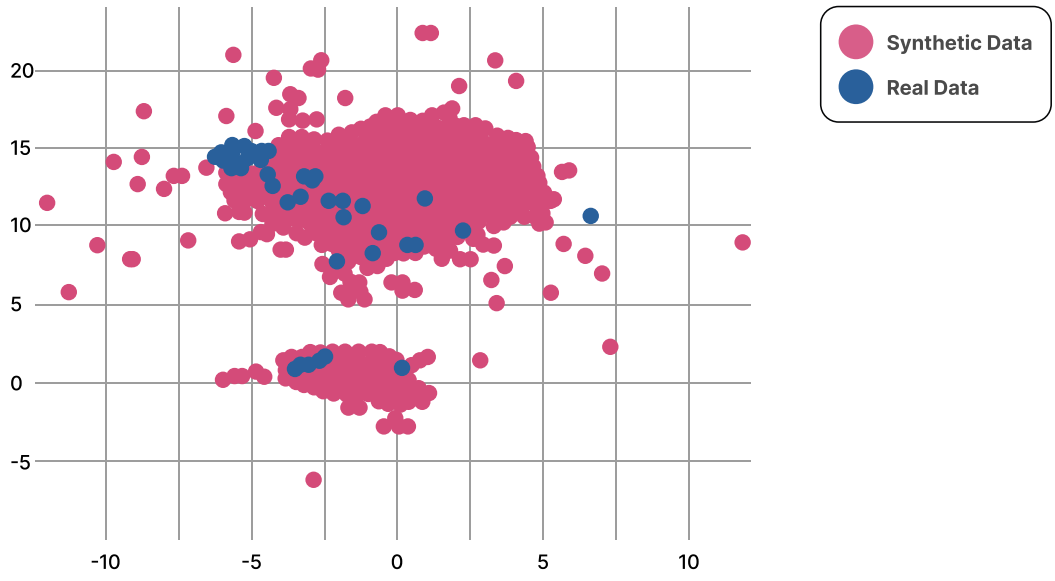


Fig. 5 Information Richness created using completely random variation.

Conversely, Fig. 5 illustrates the effect of using completely random data addition to generate novel synthetic data. Here, the Information Richness is high, but is completely untethered from Indistinguishability, providing novel but less relevant data, and highlighting the importance that the "Three I's" have together.

Effect of Real-World Samples on Information Richness

Theoretically, Information Richness can increase until the performance of an ML model is 100% perfect. And while this is always the goal, there is a tradeoff on the relevant Information Richness that can be obtained from any specific real-world sample and the cost of collecting such data. As we add relevant data points to a large set, the uniqueness of each additional point eventually begins to decrease. But as we noted in the previous post on Indistinguishability, a Digital Twin does not need to be 100% Indistinguishable to be highly useful. Similarly, Information Richness does not need to increase indefinitely.

So, how do we know when a synthetic sample is sufficiently Information Rich? We can judge this by running a sum of the uniqueness of all individual points – the sum will keep appreciably increasing if the new points are adding new, novel information. Once additional points are no longer adding new information, the sum starts to hit a plateau and we know that adding more points is no longer increasing the Information Richness of our set (Fig. 6). This is likewise reflected in our area-based evaluation method: adding points that do not provide novel information will only add points that cover an area that has already been covered, and the total area of our cluster will not change.

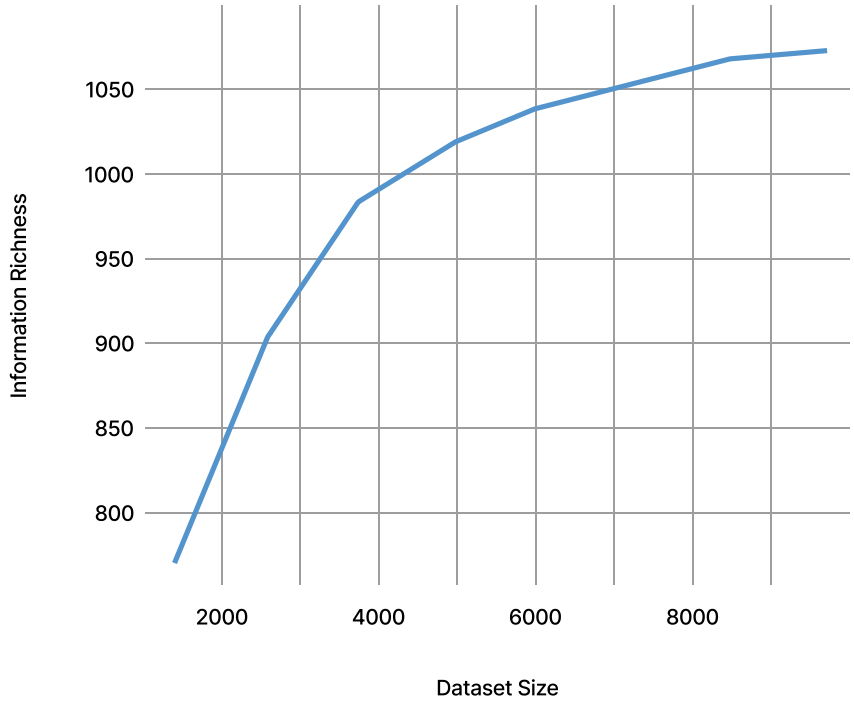


Fig. 6 Information Richness vs dataset size. The larger the dataset, the less likely it is that additional new data will be novel.

As with other aspects of synthetic data, this limitation stems from the inherent tethering of synthetic data to the real-world sample it is based on and the methodology by which it was collected. Different methodologies can yield samples that are less or more advantageous for generating greater Information Richness, and this consideration plays an important part in the last of the “Three I’s”: Intentionality.

Intentionality in Domain Randomization

Ultimately, all Information Richness, especially when produced by Domain Randomization, has value. But we believe that through a more selective and intentional use of Domain Randomization, we'll be able to create truly useful Information Richness that streamlines training and increases performance of ML models. It is this Intentionality and how we use it to shape our data sets as well as how we balance Information Richness and Indistinguishability that we will cover in the next section.

Ensuring Intentionality

So far we have discussed what it means to generate Information Rich synthetic data based on highly Indistinguishable samples generated from a Digital Twin. We have shown that we can determine how Indistinguishable our synthetic data is from the original sample and that we can quantify novelty compared to the real-world collected data. But while Indistinguishable and Information Rich data is a good baseline — the true value of our synthetic data is determined by its relevance, and ensuring this relevance requires a careful shepherding framework which we refer to as **Intentionality**.

At its core, Intentionality stems from the ever-present awareness that any ML model trained on our data must function in the nuance and chaos of the real world and not just in the customizable conditions of a simulation. And while possibilities of how to evolve Information Richness are practically boundless, they can also be operationally overwhelming. Unfortunately, a one-size-fits-all approach often results in all being fit rather poorly. Intentionality, simply put, is the tailoring of our synthetic data to the specific real-world problem itself.

How Do We Define Intentionality?

Intentionality is how we define our Domain of Operation after thoroughly understanding the Domain of Interest. With a careful consideration of the realities in which an ML model will be expected to reliably perform, Intentionality is how we choose what variations in our data to keep, to emphasize, or to prune. A Digital Twin can be used to generate very different Intentional data sets depending on what features may or may not be relevant to the specific ML model being trained. Intentionality strives for the holistic understanding of what specific Information Richness to introduce and can be viewed as the control mechanism for shaping novel synthetic data away from the distribution of the real-world sample and towards one more advantageous to robust training.

We mentioned in the previous post that Information Richness that expands the domain is also inherently more speculative. Intentionality functions as a fulcrum to balance this speculation against Indistinguishability; keeping our variations as realistic as possible, while allowing for useful aberrations (such as higher representation of edge cases than in the real-world sample) to persist.

Intentionality is the primary guiding principle of how we think about the relevance of synthetic data generation, and it shapes every step of our process.

How Does Intentionality Work In Practice?

We begin with a deep dive into a use case. To learn all about a domain, we rely on experts and studies in the field, the customer's expertise, as well as our own research. As we build our understanding, we begin to define our domain, and key criteria start to emerge. This leads to examining which conditions are pertinent and if they are variable or static. Which contexts and environments are to be expected, and which are irrelevant. What edge cases are significant, and which ones can be ignored. This process is frequently iterative and persists until we sufficiently delineate the features vital to our data set.

For an example of the variety of conditions, consider an ML model that may be used to sort inventory in a retail supply chain. While this appears to be a well defined task, significantly different data could be needed if the supply chain is for a large department store versus a highly specialized boutique. The department store incorporates a greater variety of objects, with higher number of permutations, in more complex and variable contexts all while being more likely to experience more frequent inventory fluctuations and turnovers. The patterns of item stocking, the types of changes and varieties of human error may also be different between these two environments, necessitating a further emphasis on different types of edge cases. An Intentional data set should integrate these nuances to the best of our ability, avoiding the confounding issues often caused by less-relevant information.

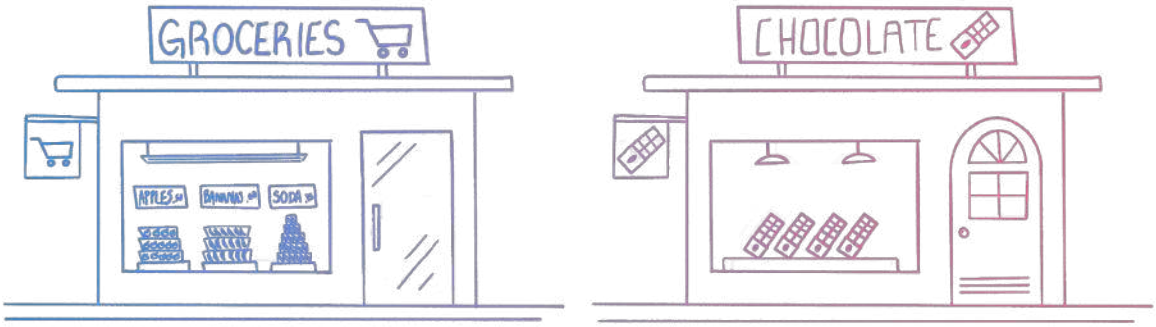


Fig. 1 Even though both businesses carry chocolate, the supply line context of the grocery store is more complex, and requires a more diverse data set for successful training.

Types of ML tasks similarly drive fundamental choices about the scope of any data set. For example, a Classification scenario will have different synthetic data needs than an Object Detection scenario. Any form of Classification model seeks to identify the presence of an already familiar object in a novel image. This means that the training and testing data should be photorealistic and generally averse to variations that stray from these criteria (indicating a more narrow domain of interest). Alternatively, a Detection scenario requires recognizing unknown and novel objects in unpredictable conditions and, thus, features a broader domain of interest (Fig. 2).

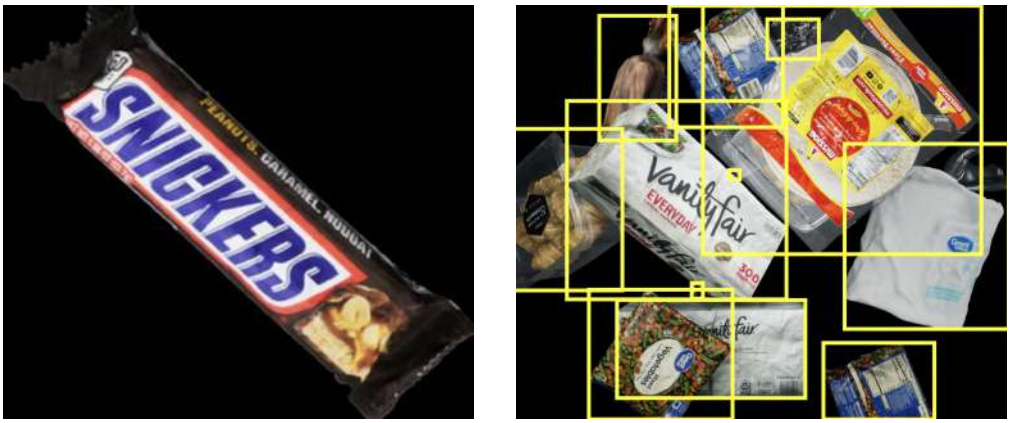


Fig. 2 A Classification scenario presents a more narrow domain (left) than that of an Object Detection scenario (right)

This broader domain of interest in part arises from the high unpredictability of the task, but it also incorporates the higher utility of non-photorealistic images for a Detection Model (data simulating input from various kinds of sensors, random variations, etc.). These are just two simplified examples of possible ML models and they will likely change with time. What is truly essential for good Intentional data is the process of integrating the understanding of any given task into the optimal data set created for that scenario.

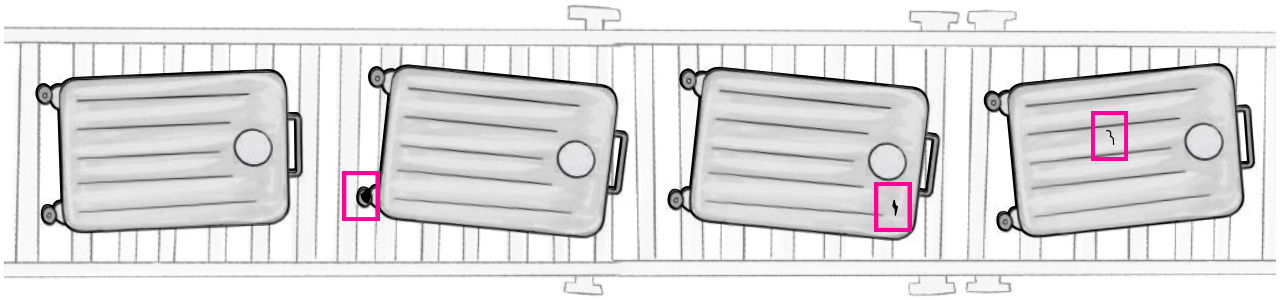
Concurrently with understanding our domain of interest, we build a Digital Twin using a large set of real-world images. Once data generated from this Digital Twin passes the Indistinguishability test, we are ready to Intentionally detach from the real-world sample using all the considerations we enumerated above. In generating this Information Richness, we may find some parallels to Domain Randomization (DR), where we are also looking for randomized variety. However, unlike standard DR, we are looking to create a bounded variety in very specific variables. This is where we draw on our study of the use case to identify points of high variation that may not be captured in the real-world sample. This may be as common as moving away from specific lights or camera lenses, to less predictable features such as variable backgrounds or objects.

What emerges is a large synthetic data set, Intentionally tailored to the scenario and the specific functions occurring in that scenario. This means that the training of this model will maximize exposure to error causing phenomena, and minimize instances that have no bearing on, or are irrelevant to, its performance.

A (Suit) Case Study

Let's revisit our suitcase illustrations from the previous sections.
For simplicity, we can imagine two scenarios:

Scenario (A) where we need to detect defects on newly manufactured suitcases and **Scenario (B)** where we need to sort customers' own suitcases for shipping or transportation.

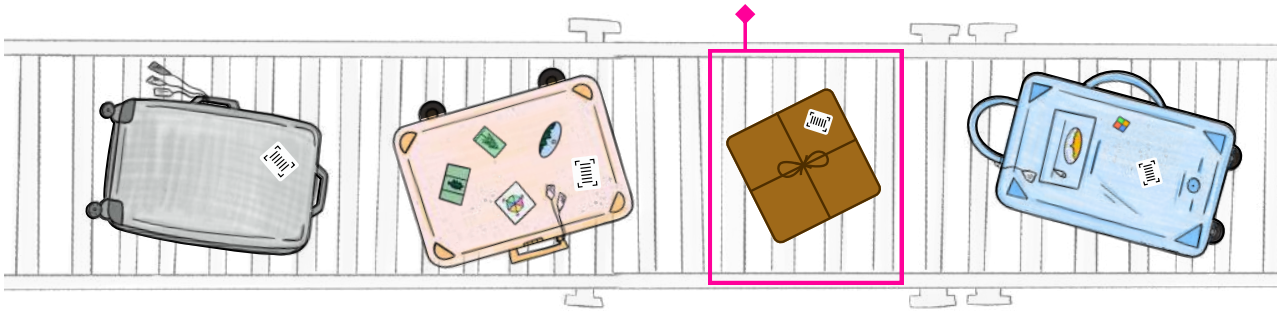


Defect Detection

Fig. 3 The narrow scope of the defect detection task is reflected in the smaller, more specific Domain of Operation necessary for this task. A good understanding of possible defects lessens the reliance on randomized data.

In the manufacturing defect detection scenario, we already know what our perfect suitcases should look like, and we can learn common patterns of manufacturing errors from the customer. We also know that we have a relatively predictable production line environment, and a comparatively low amount of variability in the non-defective products. With good quality real-world images, we can be confident of generating highly Indistinguishable data. As we evolve this data to be more Information Rich, a domain with such a defined scope, allows us to set up "guarantees", or statistical rules for how often various events are expected to occur in our data. In this case, the guarantees would include the defects as we expect to find them in the wild, as well as an over representation of those defects that can be hard to spot.

As we focus on the oversampling and higher variation of images that show defects in the suitcases, as well as images that can be falsely identified as defective (due to artifacts of lighting, lens aberrations, etc.) — what emerges is a relatively small domain of interest, and the distribution of these images should not deviate too greatly from the real-world sample.



Shipping-sorting

Fig. 4 In the shipping-sorting scenario an ML model needs to identify suitcases (and reject non-suitcases - as shown by the flagged box) as well as their various properties. This presents a much larger, harder to define domain, requiring a significantly more varied and randomized data set.

In the shipping-sorting scenario, we will run into a much greater variety of situations that are not likely to be captured in a real-world sample. In fact, the variety of suitcase sizes, forms, colors, and states of wear is so great that we do not even know what the domain really looks like. Unlike the previous scenario, there are no guarantees that we could assign here since the scope of possibility tells us that we do not even know what we do not know about this domain. To help us bridge this Domain Gap, we need a significantly more Information Rich data set, one that will drift increasingly further from the real-world sample. We need to include a bounded but large variety of randomized images, along with non-photorealistic results to help hone any identification schemas.

We will also need to include shipping labels with an emphasis on their variations. If this sorter is expected to run in different facilities, we will need greater variety in our backgrounds and lighting conditions. As we enumerate all these options, it quickly becomes evident how this data set will evolve in a significantly different direction than the defect detection data set, encompassing a much greater domain.

Even though both of these hypothetical scenarios deal with suitcases on a conveyor belt, Intentionality guides us to create two very different synthetic data sets. One is narrow and targeted, concerned with only relevant aberrations on a suitcase, while the other is broad and varied, aimed at fundamentally describing what a suitcase is. There is a vast gap between these scenarios and their synthetic data needs, and Intentionality helps us design the most efficient option for each one.

The Three I's Together

As we move into the future, the ubiquity of ML models is becoming an undeniable reality. And while we will increasingly use and experience these models in the real world, the comprehensive training they require can only happen in a synthetic one. The quality of these synthetic worlds, and the data they yield will always predict how reliably well we can train our present and future models.

Indistinguishable. Information Rich. Intentional. These three benchmark descriptors are the essential criteria by which we quantify the realism and relevance of our Digital Twin generated data.

High Indistinguishability roots our data in the reality of a concrete and observed scenario, while Information Richness allows us to mindfully broaden the horizons of our data beyond what was strictly observed. The linchpin of it all is Intentionality, directing the evolution of our data, shaping it to reflect the realities and requirements of any specific scenario – concretely defining our Domain of Operation.

“The Three I’s” together comprise a deeply interwoven methodology that helps us balance all the aspects of evolving, complex data sets — keeping them as realistic as possible, while optimizing them for each novel situation, and ultimately, decreasing the Domain Gap. This approach also helps make novel synthetic data future-proof since data that follows “The Three I’s” framework is not tied to any specific ML model, but is tailored to the use case and domain it represents. As long as this data is of a high quality, any future ML model can benefit from it. And since we are sourcing our data from an evergreen and ever-growing library of Digital Twins, the possibilities of and scope of what we can capture will only increase.

The data diets of our ML models matter progressively more, and we need a reliable framework for assessing the quality of synthetic data that in turn will yield impactful and predictable return on data investment. We believe that “The Three I’s” framework is a valuable recipe for combating junk data, ensuring relevant training for future ML models, and bringing us significantly closer to unlocking the true potential of synthetic data.

Contact Duality today sales@duality.ai