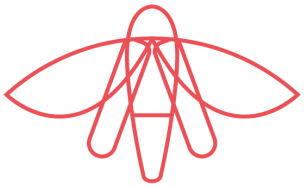


The Future of Data is Human®

Data Quality Authenticated

At Peroptyx, our mission is to design and deliver the most relevant training data and model evaluation solutions for machine learning, built around secure connections to the most reliable sources of domain-expert knowledge and human insight.

Arthur Samuel from IBM, a pioneer in the field of computer gaming and artificial intelligence, coined the term 'Machine Learning' in 1959 and it has taken from then until the last decade for its *potential* - as envisaged in the 50's and 60's - to catch up with and match its *promise*. Today, Machine Learning (ML) is being applied in new and exciting ways across technologies and sectors; from smart voice-assistants to personalised search results, and from self-driving cars to personalized financial services.

Technical dimensions of Machine Learning coupled with abstract language around its deployment and maintenance can make it difficult for business leaders to accurately assess the return on an AI/ML investment in their business, and how it will deliver a sustainable competitive advantage.

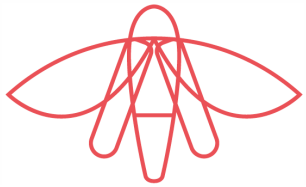
Whether a business is Digital First or undergoing a Digital Transformation, any return on investment assessment requires input from the Chief Information, Technology, Data Science or Analytics Officer who decides which data platform(s), tools and people will be required to transform proprietary business data into business value. Whether the business objective is cost saving, experience creating or revenue generating, the RoI calculation effort for an AI/ML deployment – including the cost of integration with legacy systems - should not be underestimated.

From our experience of working with Google, Apple, Microsoft, Baidu, and other early ML adopters since 2005, we have gained deep insight into what differentiates a better ML deployment. Apart from clear business objectives, what today's AI leaders have in common is the view that algorithm performance begins with the quality of data used to train the model.

While modern day machine learning tends to focus on increasingly complex 'black-box' algorithms, the reality is that training data annotation and model evaluation by humans can have a significantly positive bearing on ML project success[1] in real-world business environments.

Pre-trained models and pre-labelled data are widely available - but ensuring training data quality is at the expected level of *relevance* (i.e., timely quality) for the model to perform *consistently, reliably and to expectations over time* is a different matter entirely.

[1] Output Relevance, Execution Time, Cost/Benefit, Ethics all constitute performance.



The Future of Data is Human®

"*Fundamentally, we have been unable to predict future pricing of homes to a level of accuracy that makes this a safe business to be in.*" Zillow CEO Rich Barton

US real estate company Zillow closed its iBuying business, Zillow Offers, in November 2021 after racking up $881 million in losses in 2021 alone.

The business relied on accurate predictions from its house price algorithm that should have been able to understand whether a home was undervalued and how much the price was likely to rise in the future. The causal factors leading to Zillow's systematic overpaying for properties are likely more complicated than a machine learning model performance issue.

A standout fact about vendors who sold their properties to Zillow is that *local people* would not have paid the price that Zillow did. The data used to make predictions was likely misleading and critical data fields (i.e., those in the model that determined subsequent house valuations) were not subject to sufficient scrutiny or informed oversight.

It's a lesson that we have often seen. Machine Learning deployments without independent oversight or local evaluation can lead to unexpected and potentially expensive outcomes after going into production.

Humans and Data Quality.

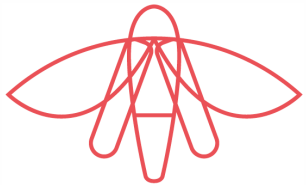
Google's 2021 paper 'Everyone wants the model work, not the data work' highlights the absence of well-defined data quality standards in ML model design and deployment.

It says: "*Data largely determines performance, fairness, robustness, safety, and scalability of AI systems…[yet] In practice, most organizations fail to create or meet any data quality standards, from under-valuing data work vis-a-vis model development.*"

The Peroptyx founding tenet 'The Future of Data is Human' was inspired by our experience working on Google search relevance from 2006 and is nicely summarized in the paper…

> "*AI model developers lack the skills and context for ensuring quality data. Because they depend critically on good data, effective AI models require* ***more human involvement*** *from both official data workers (database engineers, data curators) and from people for whom data collection and expertise isn't their day jobs but a chore (nurses, forest rangers, oil field workers and business experts intimately familiar with their data).*

It continues: "*This has a direct impact on people's lives and society, where…data quality carries an elevated significance in high-stakes AI due to its heightened downstream impact, impacting predictions like cancer detection, wildlife poaching, and loan allocations".*



The Future of Data is Human®

The Impact of Annotation Quality on AI.

To build a machine learning model that can interpret (or 'classify') data requires 'raw' content (known as unstructured data) to be identified to the algorithm using descriptions called 'labels' or 'annotations'. This is called 'training data'.

Training data is required so the algorithm can *'learn'*.

Done well, the annotation process will produce consistent, high-quality labels for training data which leads to consistent and reliable model performance.

Poor and inconsistent annotation is one of the major reasons why many ML projects fail.
A July 2021 article from MIT Technology Review on the failure of AI tools to successfully detect Covid 19 using medical imagery mentions that "many of the problems that were uncovered are linked to the poor quality of the data that researchers used", and crucially, "many tools were built using mislabeled data".

A February 2022 report from Stanford and Harvard Universities has looked at the impact of improvements in annotation quality of radiology reports when automating the interpretation of chest X-rays. They showed that higher quality labels had a statistically significant positive impact on their imaging AI. Broadly, they conclude that a data first approach has potential to improve performance of AI models in healthcare.

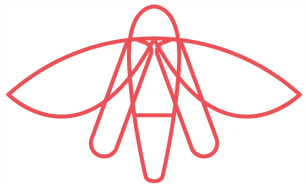
Data Centric AI.

Most organizations undertaking their AI or digital transformation journeys do not begin with the data governance and data structures required for a successful ML deployment. This partly explains the experimentation mindset of early-stage Data Science teams as they figure out how to access the structured data (e.g., databases, data warehouses) and unstructured data (images, text, video, and audio) required to build a performant ML model.

The degree of data access required is directly proportional to the degree of cultural and organizational pushback faced by data science leaders when implementing AI/ML initiatives.

Beating the 50% probability of successfully exiting this experimentation-led phase requires diligence and discipline, including the unambiguous support of organizational leadership.

After that, *80% of ML related work is data preparation*. This relates to data engineering, data cleansing and data annotation. The emphasis on data preparation is behind the emergence of a more systematic approach to ensure data quality is embedded in systems and considered important work in ML model design, deployment, and ongoing maintenance.



This is a view articulated by Andrew Ng, the pre-eminent AI thought leader, that ML ought to be more data-centric and less model-centric. Until recently, ML projects were traditionally focused on improving the model architecture, with less consideration given to the underlying training data. Ng argues for the opposite case, to focus on keeping the model 'stable' while actively iterating on the data side to drive performance improvements and value.

He gives the example of a computer vision model being deployed for identifying factory defects in steel. The model delivered accuracy levels of 76% and attempts to improve classification performance by changing the model code resulted in minimal gains.

A change of approach to a data centric one - to focus on the quality of the data by addressing inconsistencies in the training dataset, correcting noisy or conflicting dataset annotations (or labels) - resulted in the classification performance improving (from 76%) to 93.1%.

Platform-Dependent Quality.

Generalizing industry or domain specific approaches to training data quality is challenging because each ML implementation has its own unique features. The scarcity of literature on data quality illustrates the hyper-local, company-by-company approach to AI and ML implementation being undertaken thus far.

Quality control in data annotation projects is often dependent on the annotation platform itself. In other words, the application of data quality metrics is constrained by specific features (and restrictions) of a platform, instead of being guided by the underlying problem at hand.

At the same time, *outsourced data service providers* are configuring their resource supply chains exclusively around the lowest-cost geographies. They are applying quantitative and procedural methods - with mixed results – to address service defects inherent in (and inevitable with) a crowdsourcing business model. This approach is driving some of the leading annotation platform providers to move into the human resource management business to address the growing quality gap.

The Future of Data is Human®

Data Quality Authenticated.

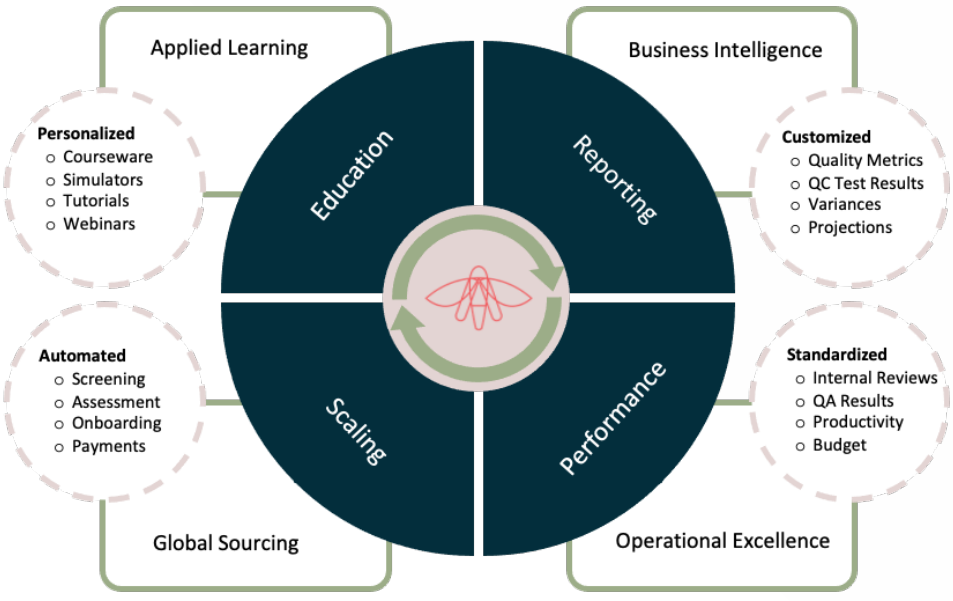
Our view is that if training data quality is a key limiting factor in reliable ML performance, then the real value for customers is in accessing data and evaluation solutions that improve ML performance for their specific implementation - without the high costs associated with building an internal team to complete these tasks.

Doing this effectively requires:

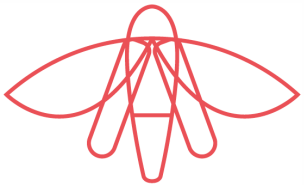
1. People who are domain-expert in their field, who have a comprehensive understanding of annotation guidelines, who have competence with a sophisticated data annotation platform, and who are retained over time for their skills and consistency.
2. A comprehensive authentication solution that persistently assures the integrity of people working on data and platforms.
3. A data driven, transparent and reliable Data Quality methodology.

Data Quality Authenticated is a framework that ensures annotator and evaluator teams deliver the highest quality ML data and human evaluation solutions for each customer use case. It is built on personalized education, transparent reporting, frequent performance feedback and the ability to scale without sacrificing quality – enabled by an industry certified platform and organization.

Data Quality Authenticated



Data Quality Authenticated represents the very best practice with insight - optimized for cost of quality, security, and customer satisfaction.



The Future of Data is Human®

In Summary

While state of the art machine learning model architectures are now generally accessible, proprietary business data used to build ML models is the most significant differentiator when it comes to ML deployments that drive long-term, sustainable competitive advantage.

Accuracy, consistency, and completeness of training data has the most direct impact on improving the performance of your AI or ML deployment.

Data Quality Authenticated represents a structured, repeatable, scalable and platform independent quality methodology that delivers ML model performance as originally intended.