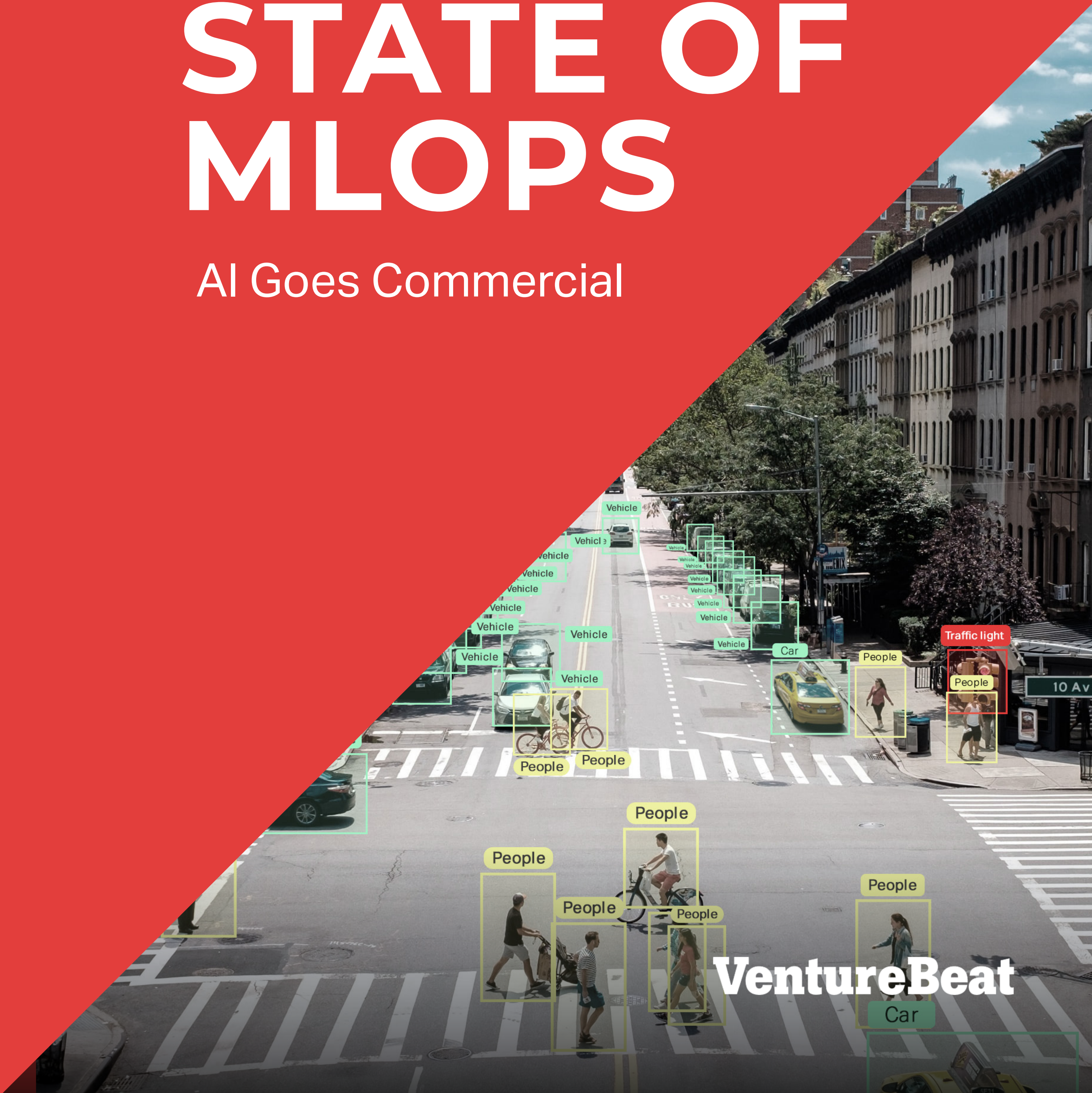
iMerit

THE 2023 STATE OF MLOPS

AI Goes Commercial



VentureBeat

The Commercialization of AI Has Just Begun

As AI moves into production, the business case for MLOps is growing

The world of AI is advancing exponentially by the day. With powerful new generative AI tools and continuous improvements in automation, both the opportunity and the necessity of having successful AI outcomes is growing. A key driver of success is building an MLOps ecosystem that enables the necessary efficiency, throughput, and oversight for AI to go into commercial production.

In the past year, an unprecedented number of commercial AI applications have made the leap from concept to launching in the real world. MLOps, the combination of model ops and data ops, has proven critical to achieving the speed and scale necessary to maximize the commercial value of AI applications. McKinsey found companies using this approach are much more likely to realize scale and value, some adding as much as 20% to their earnings.

MLOps streamlines the way AI applications are developed, deployed and optimized for ongoing value. MLDataOps - a subset of MLOps - unlocks the ability to handle data at scale through rigorous, repeatable training and testing. It ensures the sustainability of an AI project, preparing for the final critical jump to production.

AI has the potential to add $15.7 trillion to the global economy by 2030.

A PwC Global Artificial Intelligence Study shows that AI has a $15.7 trillion potential contribution to the global economy by 2030. Looking ahead, MLOps will ultimately be the differentiator in an increasingly competitive global marketplace.

To help enterprises navigate the evolving technology landscape and prepare for the future of AI production at scale, iMerit partnered with VentureBeat to gather insights from data scientists, industry leaders and professionals across industries who are on the ground, developing these products for the market.

Read on for a look at the current and future trends in machine learning, AI and MLOps.

The 7 Key Report Insights

1. **Successful AI requires better data, not more data**
   (see page 6)

2. **Data quality remains the largest obstacle to commercial AI**
   (see page 7)

3. **Human intelligence remains a central part of the AI equation**
   (see page 8)

4. **Effective, efficient data annotation is still a major challenge**
   (see page 9)

5. **Leveraging domain experts is key to success**
   (see page 10)

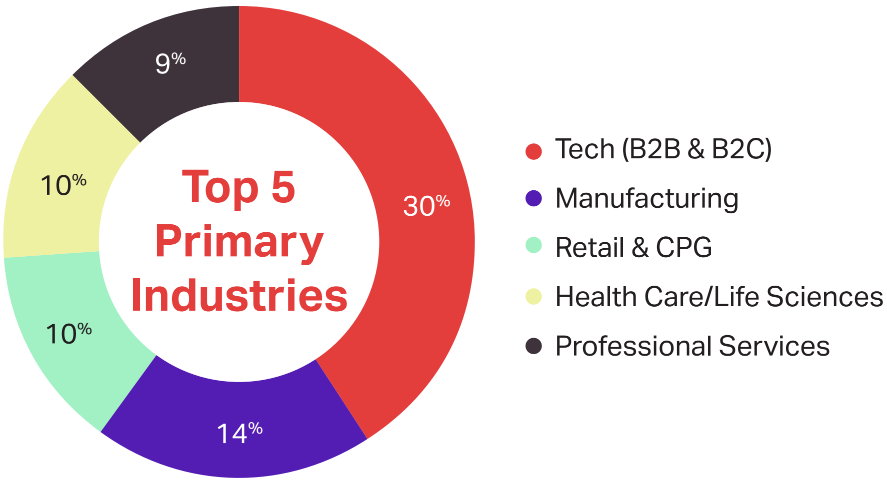
6. **Data's often the culprit for model failures – but could be a game changer instead**
   (see page 11)

7. **Solving edge cases with human intelligence is key to commercializing AI**
   (see page 12)

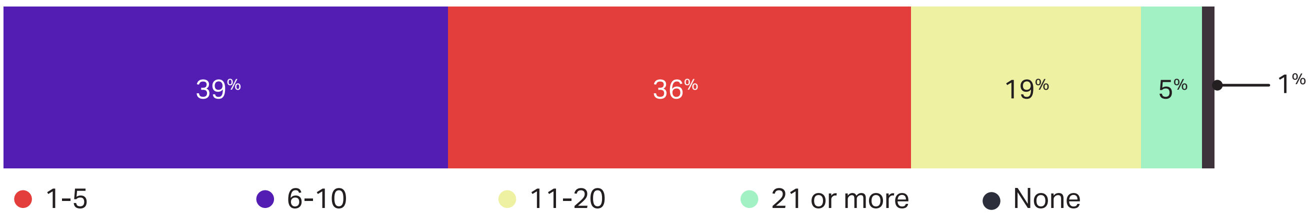
A look behind closed doors:

Who's innovating, how many projects are underway and what are the top challenges?

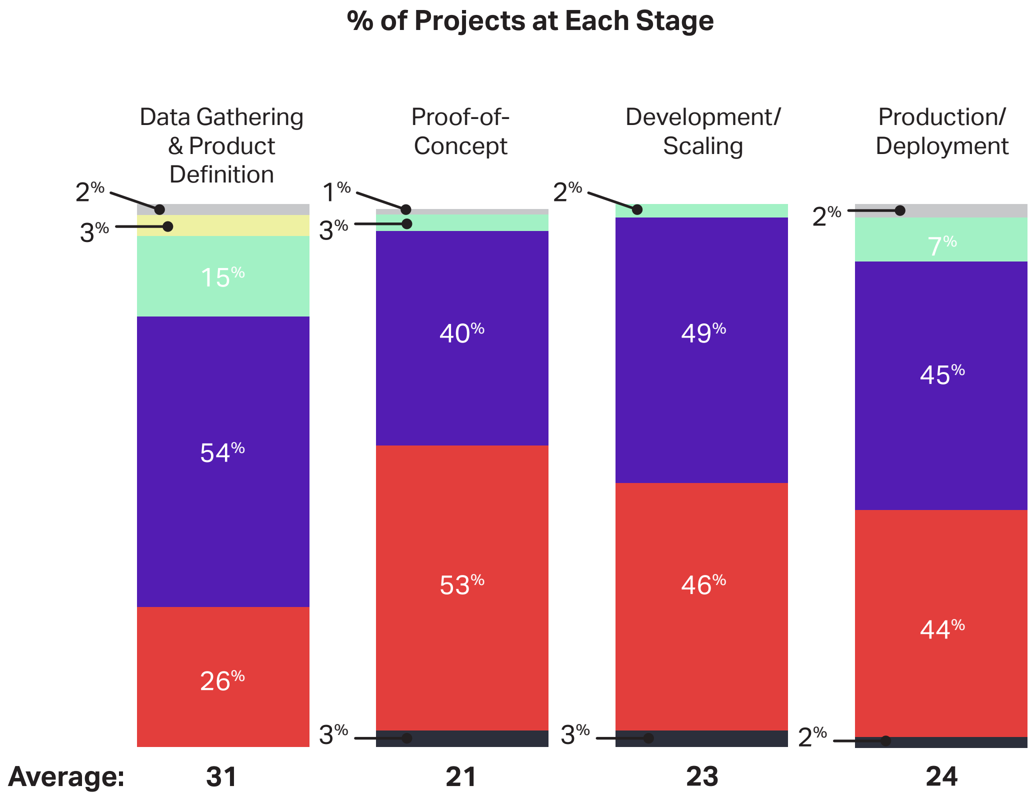
For this survey, researchers tapped ML, AI and data practitioners and leaders across industries; the top five were B2B & B2C technology, manufacturing, retail, consumer packaged goods (CPG) and healthcare.



Across industries, organizations are managing increasingly complex data pipelines, simultaneously juggling multiple initiatives, often in very different stages of development. The vast majority **(63%) have more than six ML projects in production**; **39% of companies have 6 to 10** in production; while **5% have more than 20**.



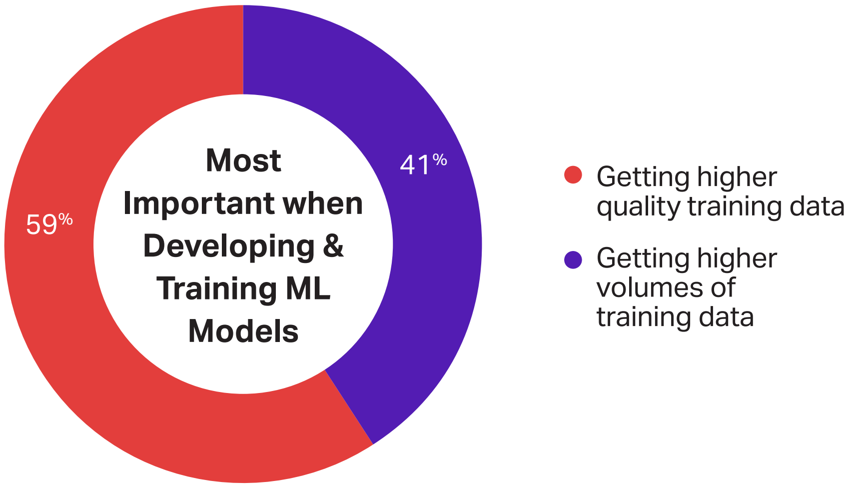
These projects span development stages: **31% are in the data gathering and product definition phase**, while **21% are in the proof-of-concept phase**, **23% in the development and scaling phase**, and **24% in the production/deployment stage.**


% of Projects at Each Stage
Data Gathering & Product Definition
Proof-of-Concept
Development/ Scaling
Production/ Deployment
2%
3%
15%
54%
26%
1%
3%
40%
53%
3%
2%
49%
46%
3%
2%
7%
45%
44%
2%
Average:
31
21
23
24


prescription transcribed
Digoxin
0.125mg
Indication
error

Insight #1: Successful AI requires better data, not more data

The significance of data quality on the accuracy of AI output has become the focus of industry conversations. Around **3 in 5 of the professionals surveyed** consider higher quality training data to be more important than higher volumes of training data for achieving the best outcomes from AI investments.



In large-scale production environments, the impact of data quality is amplified, either positively or negatively. When data is bad, the results are bad – but when it's good, it's transformative. And while it's possible to train AI using raw data (unsupervised deep machine learning), it can't surpass the results of high-quality supervised data.

Enterprise-grade AI projects often work with millions of data points, making data quality mission-critical. To achieve ROI from AI investments, it's crucial that data labeling be as accurate as possible from the start. More importantly, as AI applications move into production, ensuring AI training data is precisely labeled is critical, especially in cases like autonomous vehicles and robotic surgery, where data labeling errors could be disastrous to the long game.

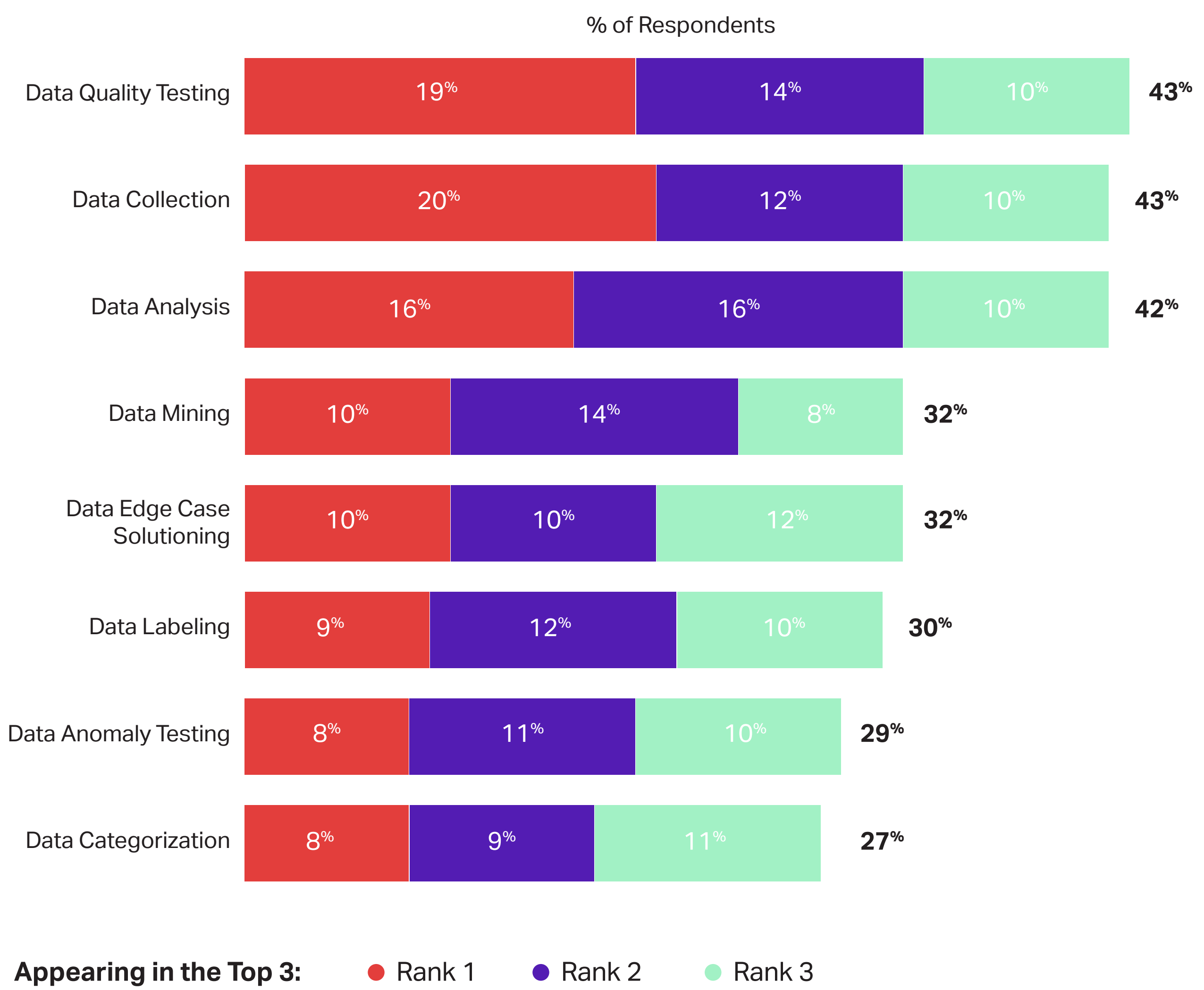
Insight #2: Data quality remains the largest obstacle to commercial AI

Experiments conducted by Andrew Ng and others show improving the quality and consistency of data being used can have an exponential impact, rather than relying on the model alone. But data has exploded — not just in pure volume, but in the amount of data necessary to synchronize for insights (for instance, multi-sensor fusion solutions for autonomous vehicle applications). Unsurprisingly, that makes data quality an urgent challenge.

The report found that data quality, expertise and volume are **the top challenges** to ML production, with **nearly half of participants** pointing to lack of data quality or precision as the number-one reason for failed ML projects, followed by a lack of expertise.

The #1 reason for failure is lack of quality data.

Top Three Challenges in ML Data Operations, Ranked



Insight #3: Human intelligence remains a central part of the AI equation

Human intervention is critical at every stage of the MLOps and MLDataOps process, from setting a goal to designing the algorithm, to ensure it's fed high-quality data. A meticulous data labeling process is the foundation of data quality, and directly impacts model predictions. In fact, well-labeled data significantly improves model performance, bumping it from an average of 60 – 70% accuracy to the 95% accuracy range.

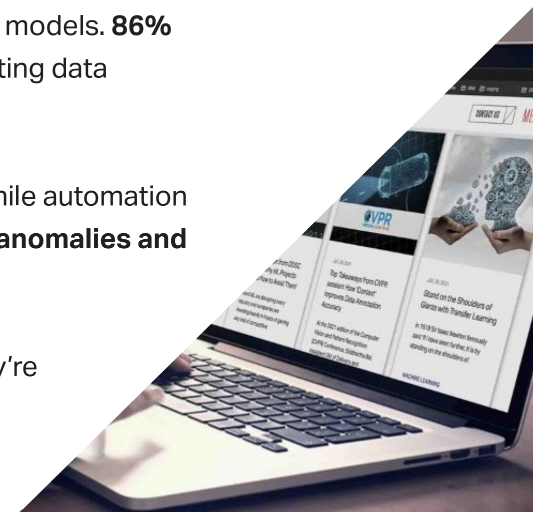
96% agreed human intelligence is key to their efforts.

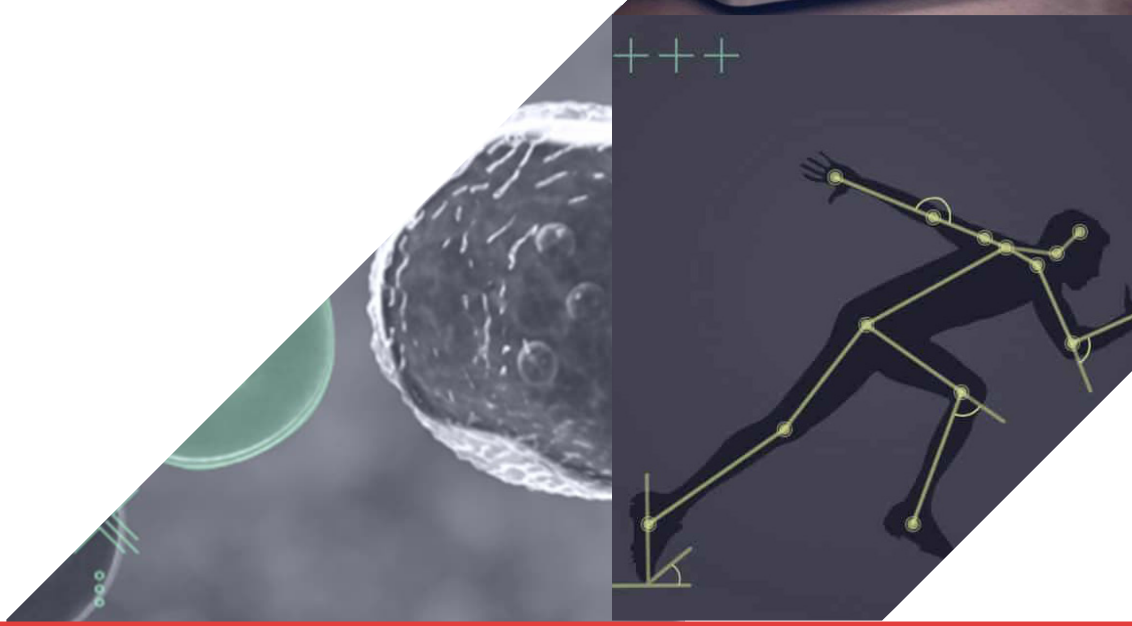
Human intervention is especially essential when accuracy is a high priority, because while automated machine learning processes may be faster, the margin for error can increase substantially. On average, **42% of all automated data labeling** requires human correction or intervention.

It's not surprising that the **vast majority surveyed (96%) agreed** human intelligence is key to their efforts, with **96% saying** human labeling is important to the success of their ML/AI data models. **86% call it essential**, and currently leverage human labeling at scale within their existing data labeling pipeline.

68% rely on a combination of both automated and human labeling, because while automation offers speed, **humans are indispensable in order to validate results, identify anomalies and offer a greater level of accuracy and quality.**

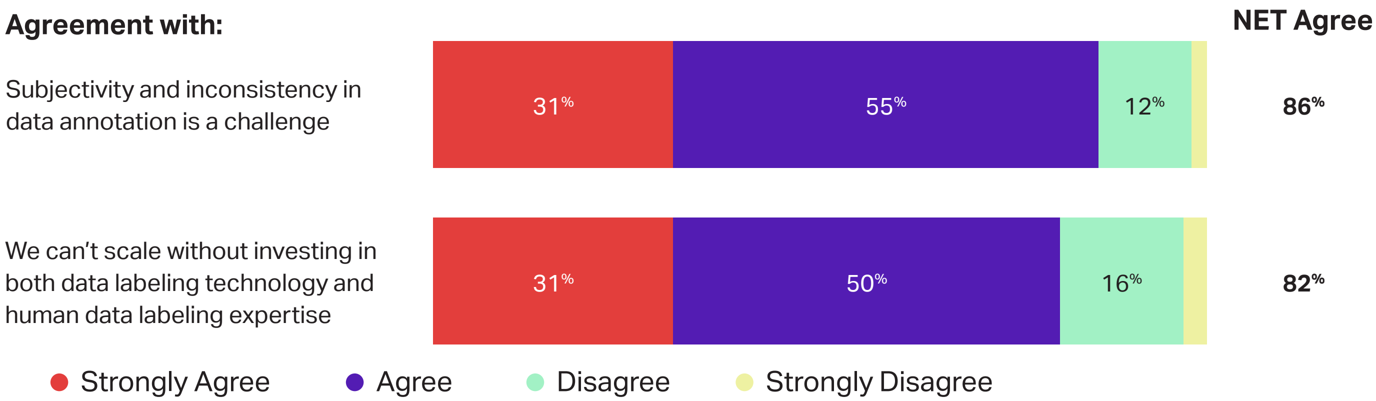
But even those who don't currently use human expertise (perhaps because they're not yet in full-scale production) consider it necessary, with **82% saying** it's important or extremely important to the success of ML models.

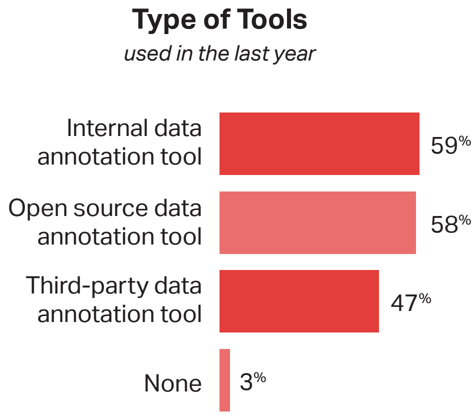




Insight #4: Effective, efficient data annotation is still a major challenge

82% of data scientists said data annotation requirements are becoming increasingly complex. This is especially true as edge cases come to the forefront (see page 12). Edge cases appear in response to the complexity and sheer variations in the real world, and need to be accurately represented in the data.





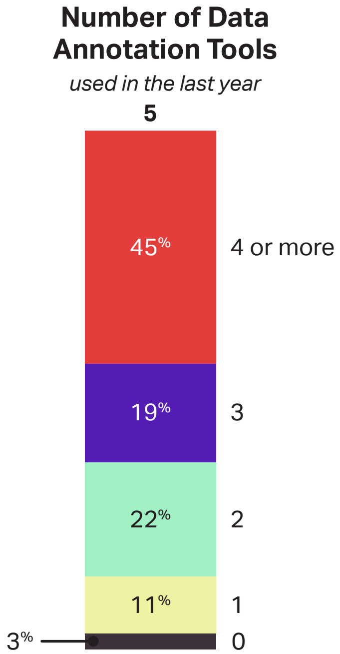
86% said the primary challenges come from subjectivity and inconsistency, often arising from unrealized gaps in their data labeling requirements. To ensure machine learning models are being trained with the right data, it's critical to define those requirements and identify the necessary domain knowledge.

To that end, 82% of professionals agree that scaling wouldn't be possible without investing in both automated annotation technology and human data labeling expertise, which overcomes these obstacles.

However, identifying the right tool for automated data annotation is another major challenge.

78% of respondents said the biggest obstacles are finding the solutions that have the right features, or are robust enough to handle their data labeling requirements—which is why **45% of companies** in the last 12 months have used **four or more data annotation tools/platforms**.

The tooling industry is perhaps not yet mature enough to offer a robust solution that covers their increasingly complex needs. Stepping into the gap, data labeling experts have become crucial to creating the high-quality data required for ML.



Insight #5: Leveraging domain experts is key to success

As more enterprises achieve full-scale commercialization, the demand for human labeling expertise of datasets will grow significantly. But not all kinds of human data labeling is considered equal.

Human data labeling is best when leveraging experts in data labeling and annotation, offering several years' experience in tagging data and understanding the requirements of different machine learning models.

This is perhaps why nearly **two-thirds (65%)** of professionals rely on a dedicated workforce with domain expertise for human labeling, compared to using crowdsourcing or freelance/contractors.

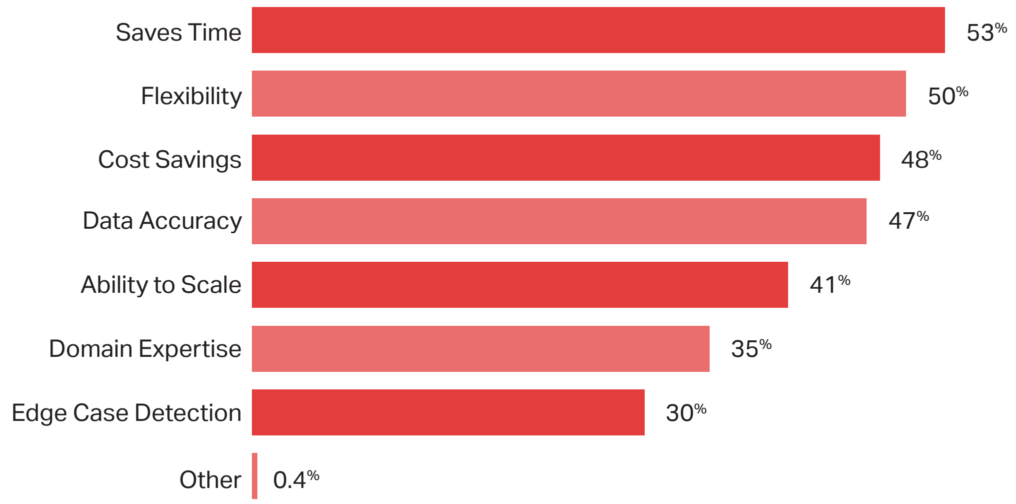
Additionally, **82% of data scientists** said they cannot scale without investing in both data labeling technology and human data labeling expertise, which makes human intelligence a critical factor for commercializing AI.

Over half of companies (55%) that are not outsourcing data labeling cite lack of data quality as the top reason their ML projects failed.

91% in the survey say they rely on outsourcing because human labeling expertise and annotation is scalable. Nearly half of all respondents said **outsourcing saved time (53%)**, **adds flexibility (50%)**, **saves costs (48%)**, and **boosts data accuracy (47%)**.

Interestingly, **over half of companies (55%)** that are not outsourcing data labeling cite lack of data quality as the top reason their ML projects failed.

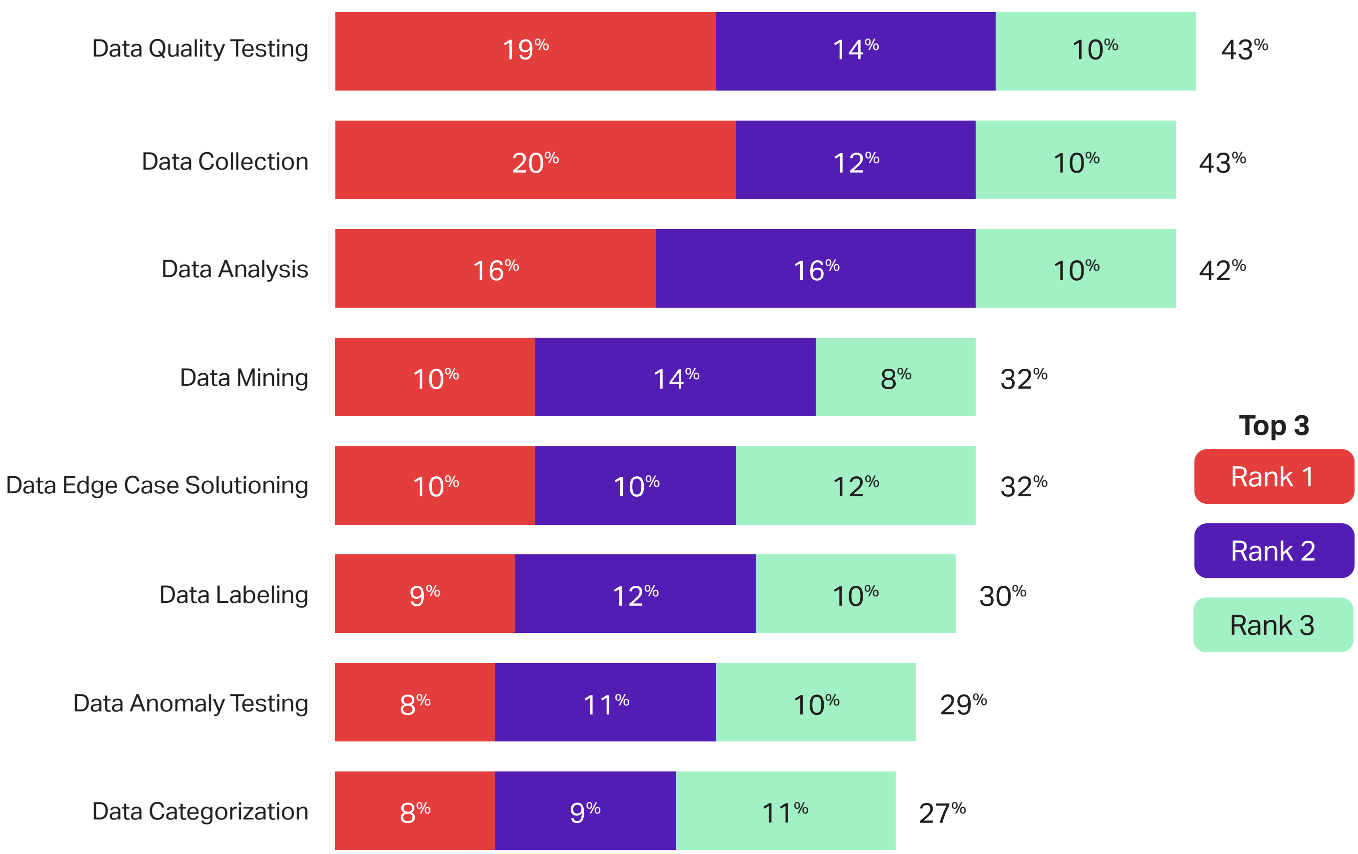
Reasons for Outsourcing Data Labeling
Among those who outsource data labeling



Insight #6: Data's often the culprit for model failures – but could be a game changer instead

Among eight different factors, AI/ML professionals ranked **data quality testing**, **data collection** and **data analysis as their top three challenges**. At the same time, when evaluating the reason for the failure of ML projects, **almost half of professionals (46%) said lack of data quality or precision was the number-one reason, followed by a lack of expertise.**

Potential Challenges in ML Data Operations



With data analysis ranked among the top three challenges data scientists face in successfully navigating ML DataOps, organizations are leveraging a variety of data monitoring tools to try and help resolve issues. In fact, **61% of data scientists** indicated they use three or more internal monitoring tools, **46% said** they use 3 or more third-party monitoring tools and **55% indicated** they use three or more open-source monitoring tools.

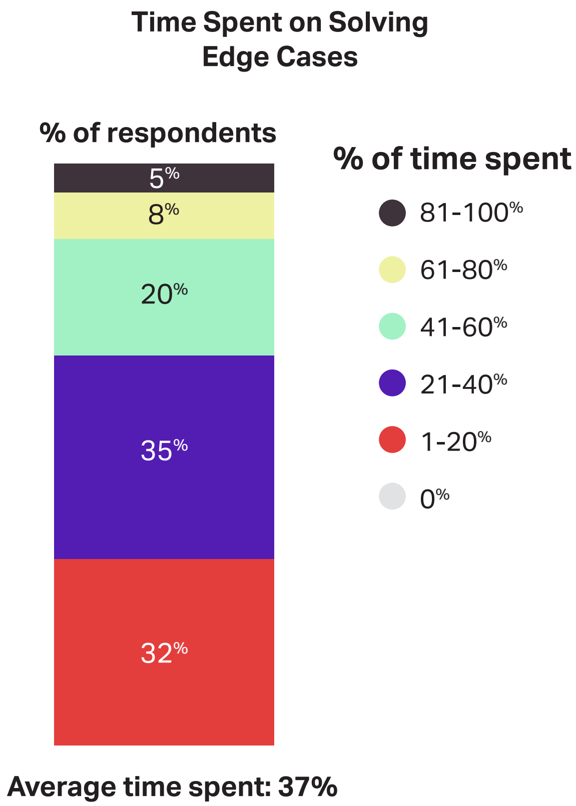
Almost half (46%) said the #1 reason for project failure was lack of data quality or precision.

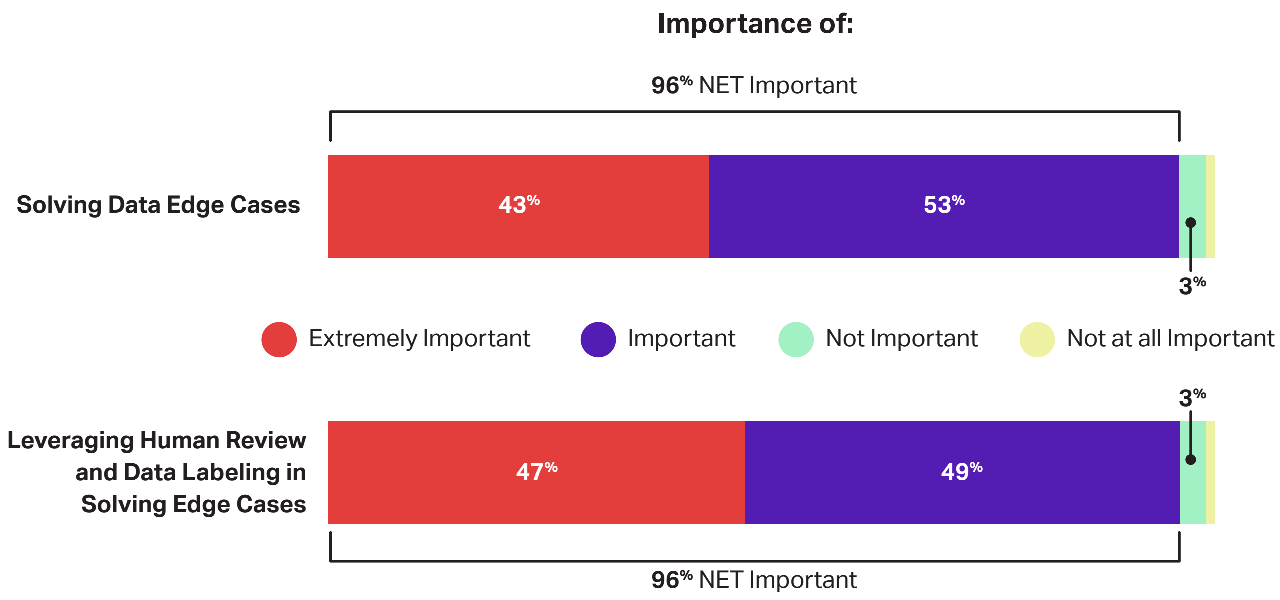
Insight #7: Solving edge cases with human intelligence is key to commercializing AI

As the number of data points in an AI data pipeline grows, so does the number of edge cases -- the rare occurrences in data that are typically seen within the last mile of the AI development lifecycle. They're the outliers caused by the complexity and sheer variations in the real world that couldn't have been predicted in a lab, and require more context and nuanced handling. **More than one-third (37%) of an AI/ML professional's time is spent working with training data to identify and solve edge cases**.

The reason edge cases are so important? Because proprietary data from edge cases is foundational to deploying commercial AI applications. Unsurprisingly, **96% said solving data edge cases is either important or extremely important, and human intelligence is central to identifying, mining and solving those cases**.

In fact, 96% also rank human review and data labeling as important or extremely important to solving edge cases.





Conclusion

Human intelligence, domain expertise and an MLDataOps infrastructure are key to commercializing AI in 2023 and beyond

MLOps, born at the intersection of DevOps, data engineering and machine learning, is setting the stage set for rapid innovation, helping bring cutting-edge AI products into production across industries like autonomous vehicles, healthcare, retail and more. Once in production, these models will need to be continually updated with new data to avoid the pitfalls of data drift and bias in their models. It becomes a fly-wheel of automation and advancement, driven by the relationship between the model and continuous high-quality data.

Human intelligence has emerged as the critical factor for commercializing AI. Every step in the virtuous cycle will continue to require human involvement, and companies will need to find ways to seamlessly integrate human expertise with tooling capabilities for auditing, monitoring and handling edge cases -- a critical step in getting ML past the final hurdles to production. It requires more than just technology alone. Going forward, the combination of technology, talent and techniques for achieving high-quality data will be the key to success.

And with AI moving faster than ever, and with so many models having already gone into production in 2022, ML data operations will ultimately decide which models are successful and which ones are not. In 2023 and beyond we'll see this play out, where edge cases will not only be differentiators among competing companies but also highlight the risks when the spotlight is put on scenarios where the AI didn't behave properly.

It's becoming increasingly clear that enterprises with the AI resources to identify the right data labeling technology and which lean on domain expertise may be the first to successfully cross the finish line.

If you'd like to learn more about AI data solutions, contact iMerit today to talk to an expert.