DefinedCrowd

Selecting Your Trusted AI Data Partner

A Step-by-Step Framework for Evaluating Training Data Providers

by **Dr. Daniela Braga**
Founder and CEO at DefinedCrowd

Introduction

More data beats better algorithms, but better data beats more data. That's according to Google's Head of Research, Peter Norvig1 and Linkedin's former Senior Data Scientist, Monica Rogati.2 We'd say their track records speak for themselves.

Industry studies support their claim, while also highlighting a substantial knowledge gap between companies just initiating AI/ML pilots and companies who have moved on to the scaling phase.

For those just starting out, "lack of high-quality training data" ranks near the bottom of their anticipated roadblocks. For organizations looking to scale their AI/ML initiatives, it ranks right near the top.

As AI technologies mature, it's becoming more and more clear every day that **data quality** is the fuel that drives models to sustained success.

However, most organizations' frameworks for selecting data providers still don't take quality into account.

We designed this quality-focused evaluation framework so that you can quit shuffling through short-term service providers, and confidently commit to a trusted data partner.

Read on to learn how to find yours.

[1] Halevy, A., Norvig, P., & Pereira, F. (2009). The Unreasonable Effectiveness of Data. IEEE Intelligent Systems, 24(2), 8-12. doi:10.1109/mis.2009.36
[2] Rogati, M. (n.d.). Lies, Damned Lies, And The Data Scientist. Lecture presented at Strata Conference, New York NY.

Index

DefinedCrowd

The Cost of Bad Data

Meet Ada, a machine learning researcher who's just started her dream job at a Fortune 100 tech company with a mandate to ramp up development on a cutting-edge Natural Language Processing (NLP) model aimed at revolutionizing product marketing.

She comes in day one with big ideas fueling the pep in her step, ready to take on the world.

However, she runs into a problem; her inherited models are underperforming. She investigates, zeroing in on the underlying training data as the primary culprit.

She digs a little deeper and finds that **20% is totally unusable.** It has to be thrown out completely. Salvaging the rest proves a Herculean task. Ada and her team end up spending more **than ¾ of their days scrubbing the tainted data,** hoping to get it into a usable state.

51%

of CIO's cite data quality as a main barrier to adopting ML technologies.[3]

80%

of an average data scientist's time is spent cleaning and collecting data.[4]

76%

of data scientists view data prep as the least enjoyable part of the job.[5]

The rest of the time, Ada is jumping through hoops as she tries to make up the lost ground. Flash-forward, and product launch looks to be almost a year behind schedule, and she's had no time to model any data.

[3] Oxford Economics, and Service Now. The Global CIO Point of View: The New Agenda for Transformative Leadership: Reimagine Business for Machine Learning. 2019.
[4] https://www.ibm.com/cloud/blog/ibm-data-catalog-data-scientists-productivity
[5] https://whatsthebigdata.com/2016/05/01/data-scientists-spend-most-of-their-time-cleaning-data/

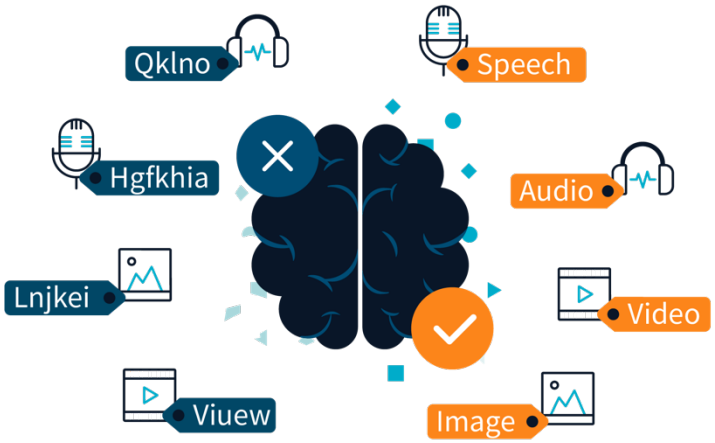
The Cost of Bad Data

Ada's story isn't a particularly happy one and could have been completely avoided had her employer simply secured a data partner capable of providing the project-specific, customized training data the project needed all along.

Unfortunately, Ada's story is all-too-common, as many companies select their training data provider based on nothing more than price-tags, not realizing the ramifications until it's far too late.

According to a 2018 study by Forrester Research, **17%** of organizations **cited a lack of "well-curated collection of data to train an AI system" as an expected challenge** in launching AI pilots. However, according to that same study, **lack of quality training data ranks as a top barrier to scaling at AI the pilot phase.**[6]

There's a well-known saying in data science: "Garbage in, garbage out." It's just a slight variation on the age-old adage, "You get what you pay for." In the world of data, it's critical to understand- at the very least- what you are paying for. Think of it as a key hire. "The lowest bidder takes the cake" is rarely a winning strategy.



[6] Oxford Economics, and Service Now. The Global CIO Point of View: The New Agenda for Transformative Leadership: Reimagine Business for Machine Learning. 2019.

The Cost of Bad Data

Data firms should face the same scrutiny as any candidate for a critical position would, with a shifted focus from their fitness as mere short-term service-providers, to their potential as long-term data partners.

They'll provide the raw material upon which your AI initiatives will run. It's just too important to do it any other way.

> **"On two occasions I have been asked, 'Pray, Mr. Babbage, if you put into the machine wrong figures, will the right answers come out?... I am not able to apprehend the confusion of ideas that could provoke such a question."**
>
> **- Charles Babbage**, Passages from the Life of a Philosopher

Deciding Who's Worth Your Time

The remainder of this paper will outline a step-by-step process for evaluating and testing prospective data firms' capabilities with an eye on their long-term viability.

The first step in that process? Knowing who's worth taking the time to talk to. Continuing with our hiring analogy, these are screener questions. Any red flags at this stage mean you can throw the "candidate" out of the pile completely.

Include these four questions on RFI's and RFP's you send to any potential data partner's way. Pay attention to how they respond. You'll be able to cross a few off your list right from the start.

1. **Do you provide SLA's for quality and throughput in contracts?**

   The answer should be yes. That means they're serious about data quality. What's more, they should be able to provide a comprehensive overview of how they track quality and which quality measurements factor into their exit criteria. Otherwise their "guarantees" don't mean a whole lot.

2. **What languages do you support?**

   It's prudent to delve into a firm's language capabilities beyond the scope of your immediate needs, particularly if your company operates on a global scale. You want to avoid the headaches, and costs, of dealing with multiple vendors spread across geographical locales. Your data partner needs to be able to handle both your current and future needs.

3. **Do you provide both self-service and premium data capabilities?**

   Self-service platforms are great for quick-turnaround jobs you can handle setting up on your own. Premium, custom data collection and annotation may take a bit longer but will play a key role in new product development.

4. **Who are your current clients and partners?**

   If they're not already working with companies you recognize, you're taking a risk.

Setting Up the Test

Knowing who's talking the talk is one thing. To find out who's capable of delivering, you'll need to put the remaining firms to the test.

In short, the way to lay the groundwork for that test, annotate a percentage of data in-house and use that "gold set" as a barometer for measuring providers' quality capabilities.

It's simplest to illustrate how this plays out with an example. However, the fundamental framework laid out here is fully adaptable to any project, any data-type (speech, text, or image), in any industry.

Scenario: Imagine a global electronics maker, we'll call them Acme Corp., has just launched their newest DSLR camera in Japan. They're looking to develop models capable of gleaning insights regarding which features people love, and where their product can be improved.

Three Steps to Set-up

1. **Source data that forms test's "base truth."**

   Using our scenario, the camera-maker web crawls 10,000 Japanese-language product reviews of varying length.

2. **Define project ontology**

   The company would define the overall desired sentiment (i.e. positive, neutral, negative) and determine specific sub-categories those sentiments should map to (i.e. shutter speed, aperture, usability and durability).

3. **Annotate 1/5 of the data in-house**

   An in-house data team of at least two people would then annotate 20% of the data collected in Step 1. This is a crucial step, and well worth the time, as these annotations will form the standard by which Acme Corp. will measure firms against one another.

Setting Up the Test

They'll test data firms on their sentiment annotation capabilities and determine their strength in Japanese-language annotation using the three set-up steps. Having gathered the raw inputs and annotated a gold set of your own, you'll have everything in place to really kick your evaluation process into full gear.

Deploying the test is simple. First, send all the raw data you collected to each prospective firm (for obvious reasons, omit your in-house annotations). As basic as it seems, how the firm handles the data transfer should be a significant part of your evaluation.

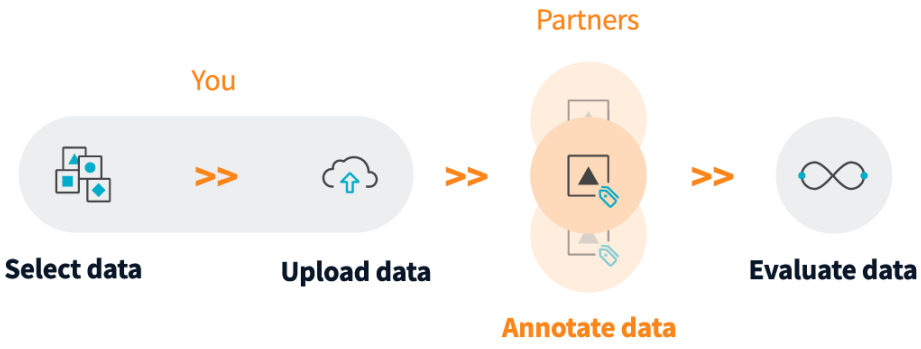
Do they have a publicly accessible API? Is uploading your data into that platform seamless? Do you run into bugs or usability issues? Are they ISO 9001 certified and GDPR compliant? These are all important questions.

To keep tracking our example, Acme Corp. would send out or upload their unstructured reviews on each platform and monitor progress until they agreed upon timeline. For semantic annotation on 10,000 textual inputs, 2-4 weeks should be enough. Make sure all firms have an equal amount of time to complete the request.

Once they've sent over their final data and accuracy metrics, simply run their annotations against your own to confirm that reported accuracy is borne out in their deliverables.

For those firms that do make the cut, compare the provided information on the remaining 80% of raw data your team did not annotate. What information is provided?

In our Acme example, to balance the subjectivity inherent to measuring sentiment, each input should have been evaluated by at least three different annotators. Inter-annotator agreement calculations for each input unit should be included.



Evaluate Results

The more transparent a firm's deliverables, the more likely it is that they truly understand the data they're delivering.

By running through these steps, you'll have a real understanding of each vendor's annotation abilities. The last piece of information you'll need? Their handling of data *collections.*

With the legwork already done, simply go back to the firms still in the running and ask for a new data collection of the least represented entities in the delivered set. You don't need to say anything more than that.

Let's say Acme Corp's web-crawled reviews have a real dearth of information regarding the camera's stabilizer, but the company knows from market research that this is a critical decision point for their Japanese customer base and included "stabilizer" as a part of their original ontology.

A data provider who knows what they're doing will be able to pick out those under-represented entities and source data that's evenly distributed throughout your desired ontology.

A firm that can do all that is worth committing to.

Congratulations, you've found a trusted data partner.

In Summary (All Steps Included)

In closing, we've outlined all steps delineated in the previous section:

Part 1: Find out who's worth talking to

What you should ask. What they should say:

Step 1: Do you provide SLA's for quality and throughput in contracts?
The answer should be yes. They should also be able to provide a comprehensive overview of how they track quality and which quality measurements factor into their exit criteria. Otherwise their "guarantees" don't mean a whole lot.

Step 2: What languages do you support?
Think beyond the scope of your immediate needs. If you operate on a global scale, you want to avoid the headaches, and costs, of dealing with multiple vendors spread across geographical locales.

Step 3: Do you provide both self-service and premium data capabilities?
Self-service platforms are great for quick-turnaround jobs you can handle setting up on your own. Premium, custom data collection and annotation may take a bit longer but will play a key role in new product development.

Step 4: Who are your current clients and partners?
If they're not already working with companies you recognize, you're taking a risk.

Part 2: Set up your test

Annotate data in-house to serve as your barometer for tracking firms' performance.

Step 1: Source data that forms test's "base truth."

You can web-crawl this data on your own, or even contract the firms to collect it for you.

In Summary (All Steps Included)

DefinedCrowd

Step 2: Define project ontology
Map out the overall ontology for your desired project.

Step 3: Annotate 1/5 of data in-house
Have two people annotate 20% of the data collected in Step 1 of this section. This is a crucial step, and well worth the time, as these annotations will form your barometer for measuring firms against one another.

Part 3: Evaluate the results

Compare the deliverables to your in-house annotations.

Step 1: Confirm accuracy
Ensure that the deliverables you receive match the reported accuracy in your own annotations.

Step 2: Compare remaining data
For firms that make the cut following the first step, compare the information provided for the remaining 80% of raw data your team did not annotate.

Step 3: Check for subjectivity
Each input should have been evaluated by at least three different annotators and inter-annotator agreement calculations should also be provided.

Part 4: The final step (collection assessment)

With the legwork already done, simply go back to the firms still in the running and ask for a new data collection of the least represented entities in the delivered set. You don't need to say anything more than that.

DefinedCrowd's first-of-its kind platform combines human intelligence and Machine-Learning powered Quality Assurance with fully customizable templates to deliver quality-guaranteed, project specific data.

Like this test? Put us through it.

Request a Trial at definedcrowd.ai
or contact us at sales@definedcrowd.com

DefinedCrowd