DefinedCrowd

Training Data Defined

Your Introduction to Training Data for AI

By **João Freitas, PhD**
CTO at DefinedCrowd

Introduction

Why Training Data Matters

Ask any given passer-by walking down the street what they know about AI and it's more than likely that "algorithms" will be among the first words you'll hear in response.

Not so long ago, you could have said the same thing about the researchers skating along the technology's cutting-edge. Indeed, for many years, research focused on the mathematical formulas driving models. Once upon a time, algorithms were king.

Nowadays though, it's data that's on top. Why? Because research led to the conclusion the Machine Learning models that achieve the best results require significant amounts of curated, labeled datasets to deliver the best results.

In 2001 Microsoft researchers Michele Banko and Eric Brill demonstrated how a multitude of completely unique algorithmic solutions to a Machine Learning problem would end up leading very different models to exactly the same results, so long as those models were all built on the same data ("Scaling to Very Very Large Corpora for Natural Language Disambiguation", Banko & Brill, 2001).

A team of researchers at Google took the argument a step further in a breakthrough paper titled "The Unreasonable Effectiveness of Data." Their most impactful finding? Expanding training data sets to include more examples was far more effective in improving model performance than any algorithmic tweaking (Halevy, Norvig & Pereira 2009).

Since then, a myriad of researchers have added one more layer of nuance to the case for data by demonstrating consistent model improvement through training data reduction aimed at weeding out any misleading or confusing examples.

Introduction

Put all that together, and a new order has been established. While algorithms may have once been king, industry is starting to realize that high-quality data sets are the kingdom, the crown, and the keys to the castle.

Or as Peter Norvig, Former Director of Research at Google (and a lead author of the aforementioned paper) puts it, "More data beats clever algorithms, but better data beats more data."

This should come as no surprise. Many Machine Learning approaches follow the fundamental tenant of "Garbage in, garbage out." It then follows that the data upon which a model is built will determine the quality of its performance. AI and ML are certainly no exception.

According to Oxford Economics, 51% of CIOs now cite "data quality" as a substantial roadblock in their efforts to adopt ML and AI products and solutions.

The good news is, more advanced industry players are finally starting to catch on. According to Oxford Economics, 51% of CIOs now cite "data quality" as a substantial roadblock in their efforts to adopt ML and AI products and solutions ("Oxford Economics, and Service Now. The Global CIO Point of View: The New Agenda for Transformative Leadership: Reimagine Business for Machine Learning", 2019).

We created this white paper for business executives to better understand the basics of training data and the interplay of data quantity and quality to ensure they can hurdle that data "roadblock" when the time comes.

Index

DefinedCrowd

The Unreasonable Importance of Data

DefinedCrowd

What is “Training Data” anyway?

Short Answer: A dataset of curated examples that unlocks a machine’s ability to understand the world.

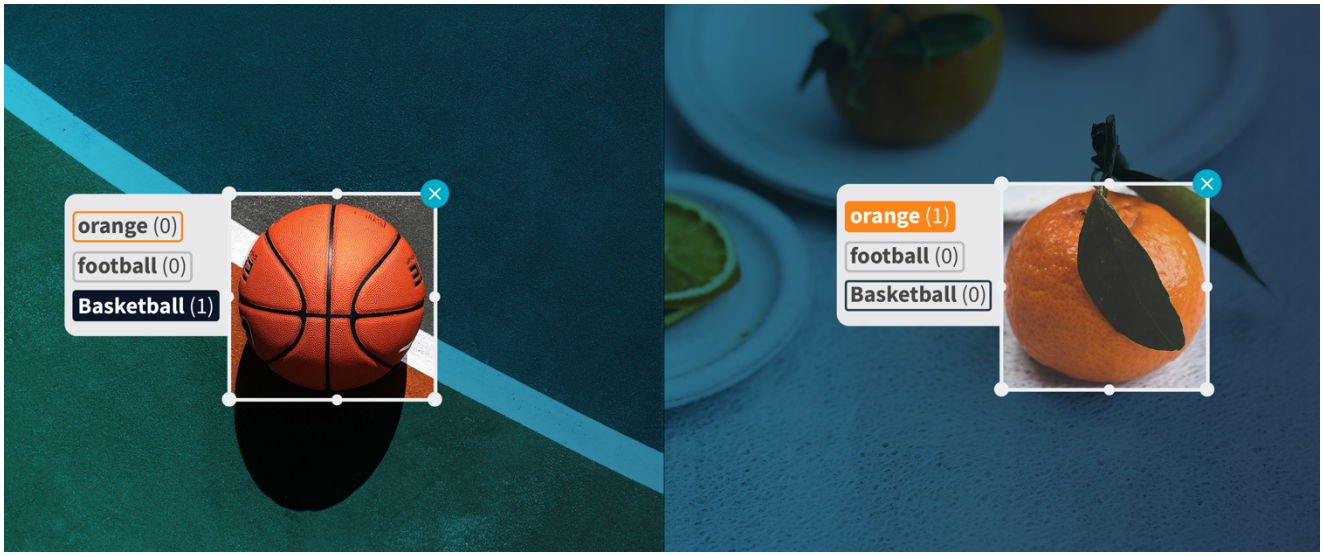
Take your pick of analogy when it comes to explaining the interplay between data and models when building intelligent machines.

If the model is the house, the data is the wood. If the model is the car, the data is the fuel. If the model is the recipe, the data is the ingredients. Think of it as the “raw material” that makes the whole thing go. As any contractor, automotive engineer, or chef would tell you, the “raw materials” require the most care and attention.

The way machines learn mirrors the way humans learn. As infants, we interact with the world, and the adults in our lives name (or “label”) things for us. “Ball,” they might say, handing us our first toy. And voila, our conceptual “model” of a “ball” begins to form.

Down the line, we encounter a basketball. It’s much larger and has many differing inform us that it too is a ball, and we update our “code” accordingly.

Later, we see another smooth, three-dimensional, round thing rolling across the kitchen counter. “Ball,” we say pointing. Here, we’re told this is, in fact, an “orange.”



The Unreasonable Importance of Data

DefinedCrowd

With enough examples, we get to the point where we intuitively understand that the object getting kicked around in the park is a "ball," but those round green fruits growing on the tree above our head are not, even if we've never seen either one of those items before.

So, how does a machine learn to distinguish a basketball from an orange? Or, parse the meaning of "Tiger got the birdie!" from "That tiger got that birdie?" Or, grasp the wide range of tonal subtext hidden in the phrase "I'm fine."

Like humans, the best-performing ML models learn through labeled examples. In this case, human annotators can play the adult's role in our analogy, providing correctly labeled data to a "developing" model so that it can learn. We call this **structured data.**

The goal? Once a model gets enough structured examples to form "ground-truth," it will then be ready to correctly interpret and act on real-world **"unstructured data"** which typically doesn't come with the benefit of labels.

80% of the world's data is stored in unstructured form, and less than 1% of it is analyzed.[5] It's common knowledge that effective ML is the key to capitalizing on all that untapped potential. We're here to tell you that correctly structured training data is the key to unlocking the best AI.

The Importance of Variability

Why Data Must Account for Variability

If the goal of AI is to process unstructured data that's too complex to be usable or meaningful, it's important to take a step back and ask what makes it so complicated in the first place.

Most of the time, variability is the culprit (invariably, you could say), a point most easily demonstrated through language and the multitudes of ways different people could express the same intent. After all, "Skip song!" vs. "Not in the mood" vs. "I hate this song," can all mean the same thing in certain contexts.

Training data collection follows the same cardinal rule as any other kind of data collection. The gathered dataset must be large enough to capture the inherent variability of the whole it is meant to represent. Otherwise, it's going to lead to unreliable results.

Most models need a large set of individual data units *and* a wide array of sources from which that data has been sourced.

For instance, you could attempt to train a speech recognition model by having someone read all the entries in a single dictionary. That's a lot of words, with little or no context.

> **"Quality is variability and variability is quality."**
>
> **- Dr. Rui Correia**, Machine Learning Engineer -

However, your model would *still* get tripped up on words used with a double meaning (e.g. "give someone the air"), informal grammatical structures, dialects, slang and jargon, and any variation in the speaker's age, gender or dialect.

The solution? More data from more people and places. That means more diverse prompt sources in order to capture informal speech and more diversity in the speakers who record those prompts. That way, the model is more widely usable as it represents a broader spectrum of speech.

The Importance of Variability

Truth is, in domains as infinitely complex as written or spoken language (for NLP and Speech) or a dynamic urban landscape (for computer vision) a dataset is never really "complete." Languages and cities are ever-changing things. There will always be gaps. Keeping models caught up takes constant collection, training, testing and fine-tuning.

However, to keep those gaps from morphing into chasms, initial datasets must account for as much of the variability a model will end up facing in the real-world as possible.

When it comes to data, size accounts for variability, making it not so much a question of quantity vs. quality but quantity as a metric of quality overall.

Or, as DefinedCrowd Machine Learning Engineer Dr. Rui Correia so eloquently put it: "Quality is variability and variability is quality." (Correia, "10 Tips For Building a Successful Chatbot", DefinedCrowd Blog, November 2018.)

Better Data Beats Bias: The Importance of Quality

DefinedCrowd

Large Amounts of Bad Data do Little Good

If variability is one of unstructured data's major challenges, and large datasets take care of variability, then that makes for a solved problem, right? We'll just keep collecting and collecting more data. Not so fast.

There are a few ways that leveraging data quantity without appropriate checks can backfire, particularly when that data originates from a limited – albeit large – set of people.

Imagine this scenario: To increase brand engagement, a company starts developing a chatbot capable of engaging in natural sounding conversations on text-based social platforms (like Twitter, Reddit, or Quora).

Let's say that company has a deep talent pool, and even deeper pockets. They build out a top-notch team to ensure the "clever algorithm" part of the equation is covered.

As for the "more data" side of things? The team comes up with an ingenious solution. Every time somebody engages in an on-platform interaction with the bot, their writing would serve as an input to further tune the model's understanding of "internet speak." Data quantity won't be an issue.

However, step back to think about the peculiarities of online discourse, particularly mob mentality and issues of anonymous dialogue and you can imagine how such a project might go off the rails.

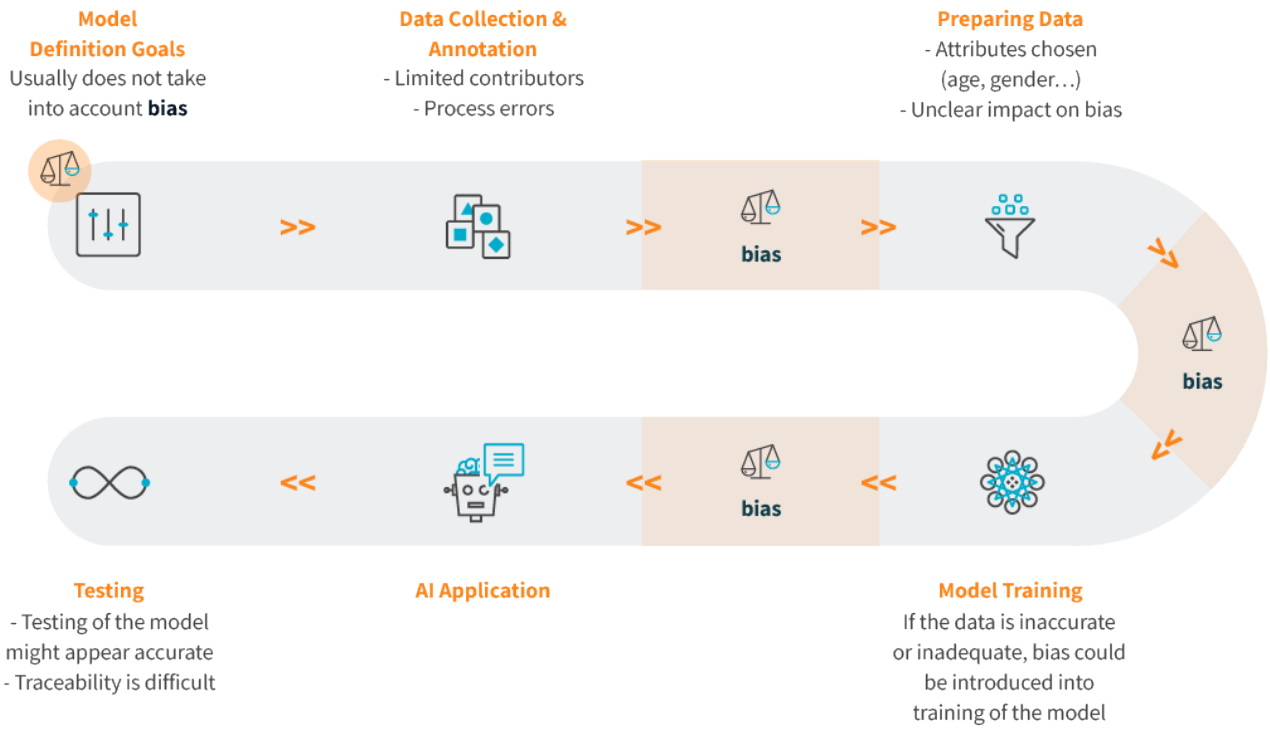
It would only take a small subgroup of users to flood the chatbot with the kinds of speech they wanted to see regurgitated – complete with personal views or biases. Even though every interaction would technically make for "more data," that data would ultimately be skewed towards the opinions and attitudes of that subgroup of users. Knowing how the internet functions, that could potentially cause a lot of brand damage, especially if the bot were to mimic offensive and antisocial prejudices.

That example illustrates a point that still plagues many first-wave, human-in-the-loop data providers who overleveraged their services on data quantity and didn't pay enough attention to quality, thus introducing bias into ML models.

Better Data Beats Bias: The Importance of Quality

The truth is, without quality guarantees to ensure that data will actually line up with a model's desired end-function, quantity alone holds limited value.

How Bias Gets into AI Models



Deep Dive: A Glossary of Useful Terms

DefinedCrowd

Unstructured Data: Information that is too complex or varied to analyze using preset fields in a database. Experts estimate that between 80-90% of the data in any organization is unstructured.

Structured Data: Information that *can* be stored and analyzed using, for example, database fields. Through machine learning techniques combined with human annotation and labeling, unstructured data can become structured data.

Human-in-the loop: An umbrella term for data structuring processes involving human annotation (most often via crowdsourcing). These processes go far beyond simply labeling data to build models. The most intelligent ML models are constantly tested and tuned with real-world usage data.

Training Data: Labeled/annotated data that forms a model's "ground truth" and its basis for understanding the world, which is why it's so critical to source large, high-quality training datasets to build functioning models.

Validation Data: Data that is used to evaluate a model's performance. Typically, teams will separate a portion of the structured data they've collected to be used exclusively as validation data, though validation data can sometimes be unlabeled data.

Real-world Usage Data: This data comes directly from users of AI products that are already in-market. For example, every time somebody asks their in-car voice assistant to call their spouse, the interaction would register as a real-world usage data point. Real-world usage data is especially useful for fine-tuning model performance, particularly in cases where the model returns an error state. Note, it is critical that users "opt in" to having their data collected in this way.

Engineered Training data (or Collected Training Data): Engineered datasets are the "bespoke" training data of AI (and DefinedCrowd's specialty). They're best for building brand new products, or brand new product features where no previous users/user datapoints exist. Data simulates product usage, often through a crowdsourced group of "users" who provide the kind of data the AI will ultimately process, and a group of expert annotators to structure that data

Deep Dive: A Glossary of Useful Terms

accordingly. DefinedCrowd's platform leads the way in Engineered Training data by allowing clients to customize the exact requirements for each group, while guaranteeing the quality of the data that comes back.

DefinedCrowd was born out of a data scientist's frustration with first-wave data providers' inability to provide the high-quality structured datasets she needed to push the technology forward.

Designed by data scientists, for data scientists, our first-of-its kind platform combines Machine Learning quality assurance procedures with cutting-edge human-in-the loop annotation practices, allowing us to be the only training data service provider in the field that includes SLA's for quality and throughput directly in our contracts.

We know that developing the future of AI and machine learning is why your researchers got into this business. Collecting, scrubbing, and validating data isn't.

We're here to help.

Just tell us what you need:
sales@definedcrowd.com

How Our Platform Works:

1. **Select your data source**

   Whether you have data to start with or not, create and configure a project with your own data parameters and rules. We take it from there.

2. **Track progress**

   Sit back, relax, and monitor progress while our platform combines human-in-the-loop annotation procedures with machine-learning powered quality assurance to deliver precision results.

3. **Receive data that's ready-to-use out-of-the-box**

   Fuel your models and reduce time-to-market with high-quality training data. From modeling, tuning, and expanding language capabilities, all the way to deployment, we've got you covered every step of the way.

DefinedCrowd's first-of-its kind platform combines human intelligence and machine-learning backed quality assurance to deliver the quality-guaranteed, project specific data necessary to successful AI initiatives.

Want to know how a quality-focused training data partner will improve your products?

We'd love to talk:
sales@definedcrowd.com

DefinedCrowd