HITL IRL: 12 Reflections on Expertise Finding and Engagement for a Large Data Curation Team

Brendan Coon[1,*]

[1]Spotify, 3 Center Plaza, Boston, MA 02108

Abstract

As ML and AI increasingly shape product development, the need for a rigorous humans-in-the-loop approach for quality control increases in importance. Impactful Data Curation teams are responsible for understanding and assessing the quality of the training data feeding into models and algorithms, and are able to package their evaluations in a consumable and actionable format. This paper covers some of the necessary steps to build a successful Data Curation team that can continuously deliver value, even as your core business or academic use case evolves. By providing an overview of what has worked during my 9 years on the team, I aim to provide an essential guide to building a new team or improve an existing one. My contention is that the unique perspective contained in this paper is advice that can help several disciplines that might be looking after a Data Curation team as part of their remit—researchers, ML engineers, product managers—get high-integrity data and algorithm evaluations from the experts they engage. Building and maintaining a Data Curation team will directly impact any product team's ability to "identify issues with usability and comprehensibility associated most closely with content quality and with the user experience." [1] It is important that you find the right people and retain them — this paper lays out how to do both. Some key takeaways the reader might acquire from this paper are how to find and identify the right experts, how to support and work with those experts, and how to retain and engage those experts. They are mostly pulled from my experience in a business environment, but can apply to an academic setting as well.

Keywords

humans in the loop, data curation, annotation, ML evaluation, subject matter expertise, curator engagement

1. Introduction

The goal of this paper is to help guide anyone working in a product development environment who needs to build or improve a Data Curation team they're responsible for.

This responsibility does not always fall on an individual as a single, dedicated task. Often the job goes to a lead researcher, ML engineer, or Product Manager despite often requiring the energy and attention of a full-time, dedicated leader who may have even been an individual contributor Data Curator themselves. This isn't necessarily the wrong organizational structure, but it can limit the amount of exposure and time the responsible party has to build and run a Data Curation team when it is only part of their remit.

This paper covers how to find the human subject matter experts, encourage retention and enable high performance - it does not go into technical details about the process of integrating data or similar experimental subjects. We know that immediate or early ML output is often wrong, unintuitive, or off-brand, and can vary wildly from end-user to end-user, but a well constructed and maintained Data Curation team can point product teams in the direction of improving that output quickly and consistently. This paper may be interesting background for those curious about how to work with a Data Curation team, but it is particularly targeted at those looking for key steps to actually find and engage the subject matter experts on a Data Curation team itself.

2. Background

In 2013, I was hired as one of the first four Data Curators at a music start-up called The Echo Nest. We worked remotely and part-time, validating data mapping via a web crawler on the order of 10k or 15k entities over several months. This project and team workflow, and others like it — experts in music and music in culture confirming computational results — proved valuable to Research and Development as they iterated on algorithms valued by multiple B2B customers. By 2015, awhile after being acquired by Spotify, the team became full-time and began branching out from label confirmation and correction to the corresponding work of heuristic evaluation. The types of work required of our team started fairly simply — evaluating one or two playlist concepts at a time over several rounds of review. But our remit eventually expanded, including but not limited to: evaluation of personalized music playlists; natural language processing (NLP) results; image quality assessment; search query fulfillment; podcast show, episode and clip recommendation analysis; track transition programming; as well as the building of a scalable taxonomy for music culture training data. Over many years, we have developed our own

Proceedings of the CIKM 2022 Workshops, 2022

*Corresponding author.

bcoon@spotify.com (B. Coon)

bespoke frameworks to package lots of nuanced analysis into actionable insights. We have collaborated weekly with music editors on discovery playlists that break up and coming artists. We have strategically shaped what should (and should not) go into music culture-centric marketing campaign data stories. This list only scratches the surface of what the Data Curation team has done in our 7 years of being full time. I have led the team since 2016, and during my leadership we have moved into the product insights part of the company, and grown from an east coast-based team of 5 to an international team of 25 subject matter experts, with some expansion yet to come over the next few years. You may have experienced some of my Spotify's personalized products, so chances are someone on my team had something to do with your experience from their role "in the loop."

3. 12 Reflections

It is possible to share much more than 12 points about how to build and maintain a Data Curation team, but I've identified these lessons as the most helpful, actionable, and applicable to a variety of Data Curation team scenarios regardless of domain.

4. Expertise Finding

4.1. Determine expertise areas

As you start building or as you inherit your team, you must determine specific areas of expertise you will absolutely need. This may sound obvious, but the way you build a team based on identified needs can impact how flexible you're able to be as your use case needs evolve. For example, when I took over hiring for my team, we were just starting to understand how we might work effectively with Natural Language Processing, and podcasts had not even been mentioned on a product roadmap yet. Once it became clear that the role of our Data Curators was going to evolve beyond "just" music expertise, adjustments were made to the hiring process to attract and screen for a broader pool of expertise. The benefit of this has been that while we maintain a core group of music experts, we are also able to provide value for the company's increasing scope. If your company's mission is made up of multiple verticals, think of the team you're building as a platform to share and serve the workload for that growth. Otherwise you can end up with several islands of Data Curators spread out due to institutional history not intentional alignment, and those teams might miss the opportunity to share knowledge, tooling, or even a consistent career development framework.

4.2. Simultaneously scope those areas

At the same time, you should accept that the scope of the expertise you're able to provide to the company must always have some appropriate limits, and that you should prioritize the knowledge that will likely improve the user experience for the most end-users. For example, if you're looking for music experts, you might find a candidate who is an authority on every recording ever committed to wax cylinder by the Edison Concert Band, but that knowledge is not practically valuable in today's music streaming market. A candidate who is integrally aware of the performers featured in XXL's latest freshman class and can apply that awareness to a recommender system evaluation is arguably of more value to your business case than someone with a PhD who can identify every 78 produced by the Victor Talking Machine Company. Prioritize the expertise you need based on the market and customer base you're serving, not necessarily at the expense of the Edison Concert band fans, but within a proper balance that favors your users.

4.3. Hire from diverse backgrounds

Your strength as a Data Curation team is proportionate to the level of diversity you're able to acquire, so you should hire a diverse team to meet whatever your needs are. If you need experts in a range of cultures or languages, do not hesitate to venture outside of a particular candidate profile. Consider a multitude of different professional backgrounds — do not exclude any academic majors or previous career paths. For example, we have had very successful members of our Data Curation team with academic backgrounds from music schools, but also business, political science, statistics, theater and English. We have hired people from companies similar to ours, but also from the DJ community, education, retail, nonprofit, and real estate. The subject matter experts you are looking for are not always the most obvious candidates jumping out of your hiring pipeline, and you will find that the strength and quality of your work will benefit from being open minded about your candidate pool.

4.4. Find knowledge lovers who can leverage that knowledge

Your curators should love acquiring knowledge, doing research, and applying both in a machine learning or iterative product environment. There are extremely capable professionals who have and can develop much of the knowledge your problem space might require, but they may not be the same individuals who are able to apply their knowledge in an actionable way. Conversely, you may find stellar project managers who are efficient at organizing a task against a deadline, but simply have too

much of a domain knowledge gap to be a fit for your team. Personality types vary of course, and this isn't an obligatory requirement, but some of the ideal candidates are people who are already participating in activities like the job they're applying for in their free time. For example, if someone you are considering is already updating online assets with sources, or painstakingly curating their own music library with what are essentially track attributes, these are very promising signs. If you do not interrogate how much your potential hire appreciates research and data improvement, you may end up with an expert who does not appreciate the application of their expertise they are now professionally responsible for. Ensure that your hires can appreciate the glory in what others might find mundane.

4.5. Develop unique screening exercises

When hiring, develop smart, non-punitive screening exercises aimed at testing knowledge, as well the ability to speak fluidly about thorny concepts (e.g. music genres.) These hiring tests should simulate the work so that both the candidate and employer know what they are getting into, but they should also help to assess curiosity, detail awareness and of course domain knowledge. For example, if you envision the candidate will be largely responsible for annotating descriptions of tracks in a particular language, test their ability to complete this work for the music or culture they have already communicated is within their area of expertise, and do this right along side tracks they may be less familiar with. Even the best experts have to do work outside of their comfort zone, so you will want to see how a candidate handles what might be unfamiliar data to them, and ask how they might start their research if this was part of a real work project in their first week of employment. This will tell you a lot about what kind of learning mindset your candidate is likely to maintain, and how satisfied that learning is likely to make them.

4.6. Balance benchmarking with bespoke investigation

When developing these tests, there are two points I want to suggest you remain vigilantly aware of:

4.6.1. False Claims

It is important that the hiring process exposes exaggerated or false claims made in an candidate's application regarding their expertise, so it is critical that you tailor some interview materials to examine these bespoke claims, while also designing identical tasks every candidate must complete for proper benchmarking. For example, if a candidate states that they have expertise in hip hop, make sure to ask them about it several times, specifically.

4.6.2. Untenable Snobbery

Simultaneously, some subject matter experts can be detrimentally snobby, so you have to investigate their professional flexibility. For example - "As part of the hiring process, some editors had to make a playlist for Susan Boyle fans to prove they could pick songs that do not necessarily align with their own taste. 'Even if it is done by a super expert, it's still for a general audience,' says Jessica Suarez, a product marketing manager at Google who serves as one of Play Music's editors. 'We're trying to reach as many people as possible.'" [2] I highly recommend this sort of assessment, as any Data Curator will eventually have to annotate or evaluate data they do not personally like or find interesting with the same level of professionalism they apply to the data they are more naturally passionate and knowledgeable about.

5. Engagement

5.1. Take on imposter syndrome head on

Recognize and embrace the imposter syndrome that is often felt by subject matter experts who are part of a Data Curation team, especially those who are joining one for the first time. Working with engineers, scientists and product managers comes with a potential learning curve that can be intimidating. A Data Curator does not necessarily have to understand python, active learning concepts, or cluster analysis. Although some curators will want to learn more about these related areas, it is not part of their required skill set or how they necessarily add the most value to your use case. Nevertheless, Data Curators have often shared with me that when compared with their counterparts in engineering and other disciplines they often feel like they don't necessarily "deserve their positions." This natural feeling but misguided sentiment must be countered directly and regularly.

For example, I and the other managers on my team loudly make the point that our work enables those engineers to iterate, those scientists to test various iterations, and those product managers to judge whether or not user needs are being met. So in fact, Data Curators are the integral glue that all of those disciplines require for ground truth and quality measurement. Curators are often able to get very close to what an actual user experience is like, and their ideas about what is not working in that experience can often expose product teams to specific examples of user painpoints. If a Data Curator feels intimidated because they cannot speak authoritatively about casual inference or a similar technical concept, we try to remind them about something they do uniquely know and can

apply — like maybe knowing all nine official members of Wu-Tang Clan. This sort of knowledge — the type Data Curators often take for granted given what disciplines they are comparing themselves to - is just as valuable when doing the majority of our work (i.e., annotation and evaluation) and you must coach Data Curators to treat their own knowledge with respect and value.

5.2. Frame the work as memorable

Data Curators can be ground truth oracles for heuristic or model training data, expert tuners of algorithms, or evaluators for algorithmic output, and are quite often all three. But your Data Curators, particularly when they are just joining, don't necessarily have this context or nomenclature. To keep this simple, try to frame most of the work encompassed in this diverse set of tasks as something memorable. Your Data Curation team should eventually learn more about precision and recall and the many related topics, but it's important that they're immediately able to connect their work with how it might be effecting models and, subsequently, end users. For example, we talk about the "the 3 T's":

5.2.1. Training

Humans annotate data with labels or free text. This ground truth or "golden data" gives models high quality and high volume training data. There is more than one approach to machine learning (ML) but typically ML algorithms learn to make decisions from this training data, depending on the particular corpus(es) a use case involves. Typically this is the part of the process people are referring to when the term "humans-in-the-loop" is used.

5.2.2. Tuning

Humans tune the model in various ways, but mostly by scoring data to track things like the limiting of accurate predictions due to overfitting, edge cases a model/classifier has not seen yet, or new categories and attributes in a schema that a model needs.

5.2.3. Testing

Humans test, validate and evaluate a model by scoring its outputs, especially in places where an algorithm has low confidence about a correct judgment or high confidence about an incorrect judgment. This is usually done with test sets to make the model robust and less likely to overfit or retain biases.

5.3. Make tenets and live by them

Your Data Curation team is not just employed to do data clean up work as an afterthought—they are there to practice a tangible, measurable and integral discipline. Most legitimate disciplines have tenets, and in Data Curation you must have bold tenets. For example:

5.3.1. Tenet 1

Every user should feel like our product gets them, regardless of who they are, where they are from, where they live, or what they like.

5.3.2. Tenet 2

Global growth is dependent on understanding cultural nuances within our products.

5.3.3. Tenet 3

Personalization is not just our products — it is truly the end-to-end user journey.

5.3.4. Tenet 4

Subject matter expertise cannot be automated, and the success of our products depends on alignment with collaborative influence.

5.3.5. Tenet 5

We reject the false dichotomy of human vs. machine and embrace the necessary and powerful collaboration of that relationship.

5.4. Develop tools and make it fun

Always be willing to develop and maintain tools and best practices that are easy for Data Curators to use, based off of sound best practices from human computer interaction research. These tools should be dependable and flexible — do not just use spreadsheets for work your Data Curation team will be repeating regularly. For example, spreadsheets work fine for many tasks, but as an annotation and evaluation tool they are incomplete interfaces. In our case, we developed an internal tool that integrates with spreadsheets, but adds a number of benefits, and is self service. The tool sets up each would be spreadsheet row as a "card" (the tool is amusingly called "cardi" in tribute to one of our favorite rappers.) It can adapt to any schema, handle enriched URIs for content playback, and produce on the fly analytics to track progress or trends from an evaluation. By all measures available, investing the time in this tool tripled our productivity, because its features were sourced from its Data Curating practitioners directly. Without the right tool, either purchased or

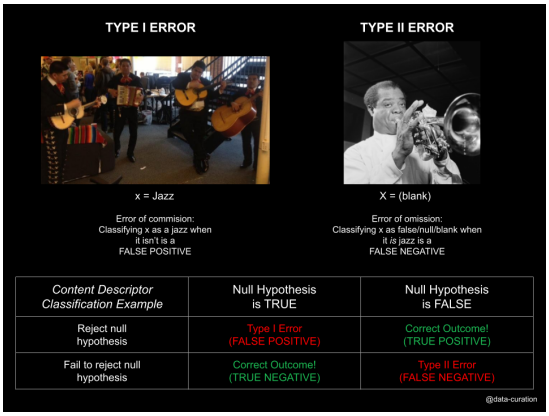

Figure 1: Design memorable ways to aid your team's understanding of statistical concepts in a manner relevant to their subject matter expertise. Here we see a fun way to remember the difference between Type 1 and Type 2 errors relevant to the domain the experts are working in. Created by the author, using photographs from his own collection and via the Library of Congress, William P. Gottlieb Collection [Public domain] (https://loc.gov/item/gottlieb.00151).

developed, you will always be leaving some time, data and quality on the table.

Also, applying bespoke best practices can be fun! There is no harm in finding relevant and creative ways to visualize important concepts germane to the work you are doing as a team as shown in Figure 1.

5.5. KPIs aren't always obvious but are always necessary

KPIs can be hard to come by and are often contextual when it comes to a Data Curation team. You can use raw counts of annotations in a database, connections made in a graph, or rates of project completion over time. Yet we have found that the better metric is something closer to the number of tests that launched over a quarter because of our team's work. When possible, any corresponding positive movement on numbers like consumption or retention is nice, but our mandate is to unlock the potential for those improvements — it is the responsibility of the product team to actually improve their code and the resulting product consumption. You can always learn a lot about how much value you are adding and where you can have the biggest impact by staying close to product development, so test launch measurement is a helpful quantification.

For example, when a product was in development, a Data Curation team "Performed a heuristic review, where (they) reviewed a number of (examples) with a variety of taste overlap scores." [1] The KPI the Data Curation cared about in the evaluation was getting something test ready by "identifying issues." This sort of focus on the value of the work proves critical to Data Curation engagement—it is the "why" the team is often looking for and can add energy to team morale and motivation.

5.6. Use the right evaluation framework

Having the right evaluation framework provides Data Curation teams with a formal and interoperable set of attributes that both focuses the feedback Data Curators generate and provides clear reporting of that feedback to stakeholders. For example, our Data Curation team has developed a "Content Recommendation Scorecard" for evaluating products or listening experiences against acceptable quality levels. Given the cognitive complexity of trying to leverage subject matter expertise in an objective way, the framework allows the team to rate a playlist or a track using several dimensions of quality - attributes like coherence or representation. When Data Curators and product teams are speaking an overlapping language, curators can ensure that they are evaluating systems consistently, and product teams can determine takeaways like "the new approach more strongly met our criteria in terms of the attributes we wanted to optimize for." [1] A detailed framework might be take time to construct and fine tune, as a healthy level of inquiry should be applied within whatever dimensions you deem appropriate. Before you develop a more rigorous evaluation framework, you can keep it simple with something like:

5.6.1. Personal Relevance

Does the recommendation match user tastes and personal preferences?

5.6.2. Cultural Relevance

Does the recommendation account for the current cultural or localized context, like contemporary trends or appropriate language?

5.6.3. Expert Artisanship

Does the recommendation feel brilliant - made by someone who knows the material inside and out and its relation to user taste?

These tasks require thoughtful work and consistent standards. Without sampling actual user segments across our most important cohorts to see and hear what various product experiences are surfacing to them, you are always sort of guessing. Data Curation removes some of that guesswork, enabling stakeholders with directional analysis that leads to beneficial action.

6. Conclusions

Some key takeaways from this paper center around how to find and identify the right experts, how to support and work with those experts, and how to keep them engaged to retain them. They are often pulled from my time in a business environment, but can also apply to an academic one. Building and maintaining a Data Curation team will directly impact any product team that leverages their expertise. Finding the right talent and engaging that talent to retain them is an important consideration, and as I have articulated in this paper, there are specific steps anyone responsible for a Data Curation team can take too optimize for both.

Acknowledgments

Thanks to my entire Data Curation team, past and current, and my colleagues in Spotify's Insights and Research communities, especially Sam Way, Claudia Huff, Aditya Ponnada, Ang Li, Praveen Ravichandran, Mounia Lalmas-Roelleke, Henriette Cramer, and Laura Lake for your guidance and support. This paper would not exist without all of your generously shared wisdom.

References

[1] J. Lamere, A look behind blend: The personalized playlist for you... and you, 2021. URL: https://engineering.atspotify.com/2021/12/a-look-behind-blend-the-personalized-playlist-for-youand-you/.

[2] V. Luckerson, These are the people picking your next internet radio song, 2015. URL: https://time.com/3947080/streaming-music-human-curators/.