THE BLACK UNICORN

CONTENT CREATION CASE STUDY DOCUMENT

Client: Lexunit
Industry: IT (AI solutions, Data Science)

Title: How to avoid falling down the Data Cascade?

Published: November 2021

BENEDEK, Gergő
gergo@theblackunicorn.eu
+36206175791

theblackunicorn.eu/

How to Avoid Falling Down the Data Cascade?

Published:

November 11, 2021

Model is King, but Data is what's flowing in his veins. An exciting study points out that project teams in the AI industry are mostly motivated to build and train interesting AI models (create "magic" with the algorithms), but managing the data that is required for training and running the model is considered more like a chore. It's not 'glamorous', and it's not getting enough attention, to the extent that it even endangers the project process itself. Many AI projects are unable to take off at all, for this very reason. That's a scary prospect for anyone working on integrating machine learning into their core business. We highlight some of the most interesting findings of the study and add some of our own experiences to it, coming from years of running AI projects of our own.

This report is based on the results of a study on practices and structural factors among 53 AI practitioners in India, the US, and East and West African countries, applying AI to high-stakes domains, such as landslide detection, suicide prevention, and cancer detection.

AI projects in the study, by industry:

Health and wellness (19) (e.g., maternal health, cancer diagnosis, mental health)

Food availability and agriculture health (10) (e.g., regenerative farming, crop illness)

Environment and climate (7) (e.g., solar energy, air pollution)

Credit and finance (7) (e.g., loans, insurance claims)

Public safety (4) (e.g., traffic violations, landslide detection, self-driving cars)

Wildlife conservation (2) (e.g., poaching and ecosystem health)

Aquaculture (2) (e.g., marine life)

Education (1) (e.g., loans, insurance claims)

Robotics (1) (e.g., physical arm sorting)

Fairness in ML (1) (e.g., representativeness)

What is a Data Cascade?

The study is focused on 'Data Cascades' - situations where an AI project goes wrong, specifically because of a data problem, not a model problem.

A 'cascade' is an event that is causing a 'technical debt' - which means the project hits an unforeseen obstacle requiring extra effort for getting the desired results, or derailing the project entirely or momentarily, or at least significantly lessening the efficiency of the solutions in use.

Staggering, but true: 92% of all the practitioners asked in the study experienced at least one event which could be identified as a Data Cascade.

Of course, these are not your everyday automation solutions, but fairly complex endeavours.

The writers identified a 'high-stake' AI project as something that is characterised by:

- high accountability

- the requirement for inter-disciplinary work

- constrained resources

You could say that in the business world, appropriate resources and expertise should be available, but the reality is that sometimes you have to make things work under less-than-ideal circumstances. We strongly believe that every business leader can learn from these examples.

The practitioners tried their best to get things done, obviously. And, because they are human beings lacking the complete perspective of their projects, they sometimes tried to make the model work even though the data was partially compromised. The authors of the study say that the practitioners 'did not appear to be equipped to recognise upstream and downstream people issues'.

Clear communication is key, protocols are essential

Quite simply, when the developers design a process, the feedback of the field workers is often not integrated early enough. The people who are responsible for the running and maintenance of the AI project itself are not trained thoroughly enough to realize the importance and value of appropriate data handling. Data becomes fuzzy and scarce, and it's dependent on the partners.

A LEXUNIT EXAMPLE: UNAVAILABLE DATASETS FORCE US TO CREATE WORKAROUND PROCESSES

THE PROJECT:

During the preparation phase of a project for a client in manufacturing, we received information that a 'Manufacturing Plan' will be available, so we can see what kind of parts are required in the factory in the next few days or weeks.

THE DATA PROBLEM:

However, we never received such a plan.

It turned out that instead, the process was done by a qualified professional on a daily basis:

1. Reviewed the priority list of tools that need to be manufactured

2. Checked what parts are available for the day

3. Cross-referenced it with the available manpower

4. Created the daily manufacture plan accordingly

Obviously, this meant that we can't start working on the optimization of the planning software because we don't have the very resource that should be fed into it: the list of required parts for the next time period.

OUR SOLUTION:

Instead of relying on a longer term manufacture plan, we built a tool that automates the process mentioned above, so the starting data is the order list itself, and our software creates the Manufacturing Plan from it.

WHAT COULD WE HAVE DONE DIFFERENTLY?

You can always go and ask again: 'Are you *sure* you've got this piece of crucial data you say you can provide?'

One even better way to do it would be to ask the client the describe the process of how exactly that piece of data is generated. You will instantly see the red flags, and you can prepare accordingly.

Data work is often ‘taken for granted’, it’s not properly recognized. The practitioners quoted in the study described it as ‘time-consuming, invisible to track, and often done under pressures to move fast due to margins—investment, constraints, and deadlines often came in the way of focusing on improving data quality’.

The study points out that it’s sometimes ‘difficult to get buy-in from clients and funders to invest in good quality data collection and annotation work, especially in price-sensitive and nascent markets like East and West African countries and India. ‘

It’s a bit like running a mill, but not caring about the quality of the grain. Which will adversely affect the quality of the flour, for sure.

Commitment cannot waver through any phases, including the data phase
The study goes further: ‘Clients expected ‘magic’ from AI—a high-performance threshold without much consideration for the underlying quality, safety, or process—which led to model performance ‘hacking’ for client demonstrations among some practitioners. ‘ We’re sure that there are many AI project managers out there who can relate - you want the project to succeed, but you also want to do it within the limitations of the expectations coming from further up the chain of command. Cutting corners can be tempting!

It seems like we are facing a systemic problem which makes AI projects so results-oriented that an unrealistic expectation is being developed by the clientside. The study points out that AI trainings and courses think of data as a 100% clean and available input material, but in real life, you never see such things as ‘clean data’.

Within IT, we've got data scientists for that, but when the input source is not fully digital, like in many of the high-stakes projects, their efforts come at a stage where it might be too late. This leads to a whole lot of AI projects going unfinished.

A LEXUNIT EXAMPLE: TYPOS DERAIL AN ALGORITHM
THE PROJECT: While optimizing a manufacturing process, we've been using data sets that describe what kind of parts are needed to build the products. THE DATA PROBLEM: One of the product descriptions contained a typo: a listed tool needed 800 pieces of a part according to the description, but in reality, it needed only 8. When the prediction algorithm started to work, it started shaping up nicely, when it suddenly spiked. OUR SOLUTION: We shut the algo down to take a look at what's going on, and found the typo. Luckily it was obvious, because not a single other tool required hundreds of pieces from the same type of part, so we managed to identify the problematic product description quickly. WHAT COULD WE HAVE DONE DIFFERENTLY? If you think that the product descriptions should be checked thoroughly before we feed it into a machine learning process, you are absolutely correct, however, it's not a trivial thing to do when your dataset contains 4000 products made out of 8000 different part types... Beyond taking a few precautions, you really need to rely on the client to deliver clean data. What might be worth doing is to emphasise the importance of that 100% 'cleanliness' and ask the client to review the chances where typos or similar errors could happen, to minimize this risk.

Causes of a cascade

So what are the main causes for data cascades? What can derail an AI project before it even gets going?

1. Physical world effects

Weather, wind, sand and dirt can mess with sensors and camera images, easily. These effects should be thoroughly analyzed before the data flow even starts, because even surprisingly small contamination can cause disturbances and lead to cascades.

We need to make sure that the physical limits and risks are thoroughly accounted for, and we have regulations in place for handling them.

A LEXUNIT EXAMPLE: MACHINE VISION DATASET CORRUPTED BY ALIEN OBJECT...
THE PROJECT: When you create training datasets from a batch of images, the content of the images is highly controlled. Even small discrepancies can vastly lower the effectiveness of the teaching process or render the algorithm completely useless. THE DATA PROBLEM: During the automated photo shoot process, someone walked into the frame… and stayed there for minutes... OUR SOLUTION: It's a bit baffling, honestly, because you would think it's obvious. When we are training a system what a vehicle is, for example, we don't want humans on the picture. But no, basically nothing is obvious. WHAT COULD WE HAVE DONE DIFFERENTLY? Clear communication, emphasis on data purity, clear instructions. Maybe they thought that standing *next to* the vehicle is not a problem, if they don't cover any surface of it. We should always do the best job possible in highlighting possible dangers and set clear rules.

2. Inadequate application-domain expertise

When the people working on the data input systems have to make decisions that are beyond their skills and knowledge, they will be forced to make guesses and they will make mistakes. Discarding, correcting, merging or a full sequence restart requires careful consideration to do effectively, which is unrealistic to expect from someone without proper preparation. This kind of problem plagued a lot of AI projects in the study, from healthcare to insurance. If the ‘ground truth’ is set incorrectly, the model will not be able to provide meaningful results.

The workers responsible for managing data entry should receive proper training or should be helped by qualified professionals, even with realtime remote guidance if necessary.

A LEXUNIT EXAMPLE: DATA SET IS CONTAMINATED BY INAPPROPRIATE HANDLING
THE PROJECT: The data set is almost always tweaked, modified, tested, in most of our projects. You need to be extra careful in having up-to-date information about how and why a certain piece of data is entered. THE DATA PROBLEM: It happened to us that the client added some values manually, for testing or compliance purposes, and simply forgot to remove the dummy data, or even mention that it was there... OUR SOLUTION: Every interaction with the dataset is logged, and the data items have several tags which they can be filtered by, so we can avoid problems like this in a live product. WHAT COULD WE HAVE DONE DIFFERENTLY? The phases of development should be clear, so the client always knows which dataset is which and what kind of interactions are allowed at the moment. Data purity should be highly emphasised from the onset. Obviously not everyone has a clear understanding of how exactly machine learning and automation works, so you need to make sure that things don't go South because someone involved in the project thinks 'it's okay' to tamper with the systems 'just a tiny bit'.

3. Conflicting reward systems

Sounds simple, but the researchers found that the tasks of data collection were simply added to the regular duties of the professionals, without them being compensated for it. It was extra work for them, for no extra pay. Obviously, data quality suffered.

4. Poor data documentation

The study found that there were several cases where data collection was subpar, and the practitioners had to make assumptions to fill the void, and sometimes discard whole datasets - in one case, 4 months of valuable medical research data.

The problem is that standards and conventions can be different between organizations, groups of professionals, and even minor differences can trigger a data cascade, so this should definitely be a focal point of every project where data gathering and analysis happens within different teams.

Conclusions

As it is written in the study: 'Our results indicate the sobering prevalence of messy, protracted, and opaque data cascades even in domains where practitioners were attuned to the importance of data quality. We need to move from current approaches that are reactive and view data as 'grunt work'. We need to move towards a proactive focus on data excellence.'

There should be clarified processes and standards to make sure data quality is sound through every phase. This requires proper infrastructure and appropriate incentives for everyone involved.

'Despite the primacy of data, novel model development is the most glamorised and celebrated work in AI—reified by the prestige of publishing new models in AI conferences, entry into AI/ML jobs and residency programs, and the pressure for startups to double up as research divisions.' Model work gets almost all the spotlight, so business leaders should not make this mistake. Carefully consider every segment of the pipeline, and do not let yourself be swayed away from allocating the right amount of resources to the data work.

This problem starts at the education level. AI literacy is almost synonymous with model development, so there aren't enough graduates skilled in the art of working with data, according to research from 2016.

There's an old meme quote in the NLP community, which unironically highlights the problems around the perception of data work: "Every time I fire a linguist, the performance of the speech recognizer goes up" . This is a prime example of 'cutting corners', generating quick value while sacrificing long-term development.

When the participants of the study have been asked about good practices, they basically mentioned nothing new or revolutionary, they just brought up tools and methods that are already used in software development, but the data side sometimes simply doesn't get that treatment:

- shared style guides

- thorough documentation

- peer review

- clearly assigned roles

All of these solutions 'compound uncertainty', and help projects avoid data contamination.

What should you do to save your own precious AI project from tumbling down a data cascade? We hope we've been able to give you some ideas and direct your attention towards the data side of the AI projects - it surely is the unsung hero which can make or break the value generation process!

The cover image is from the study itself. If you are interested and want to know more, but don't have time to read the whole 15 pages, here's a 5 minute video about the findings of the research!

***

BENEDEK, Gergő
gergo@theblackunicorn.eu
+36206175791
theblackunicorn.eu/