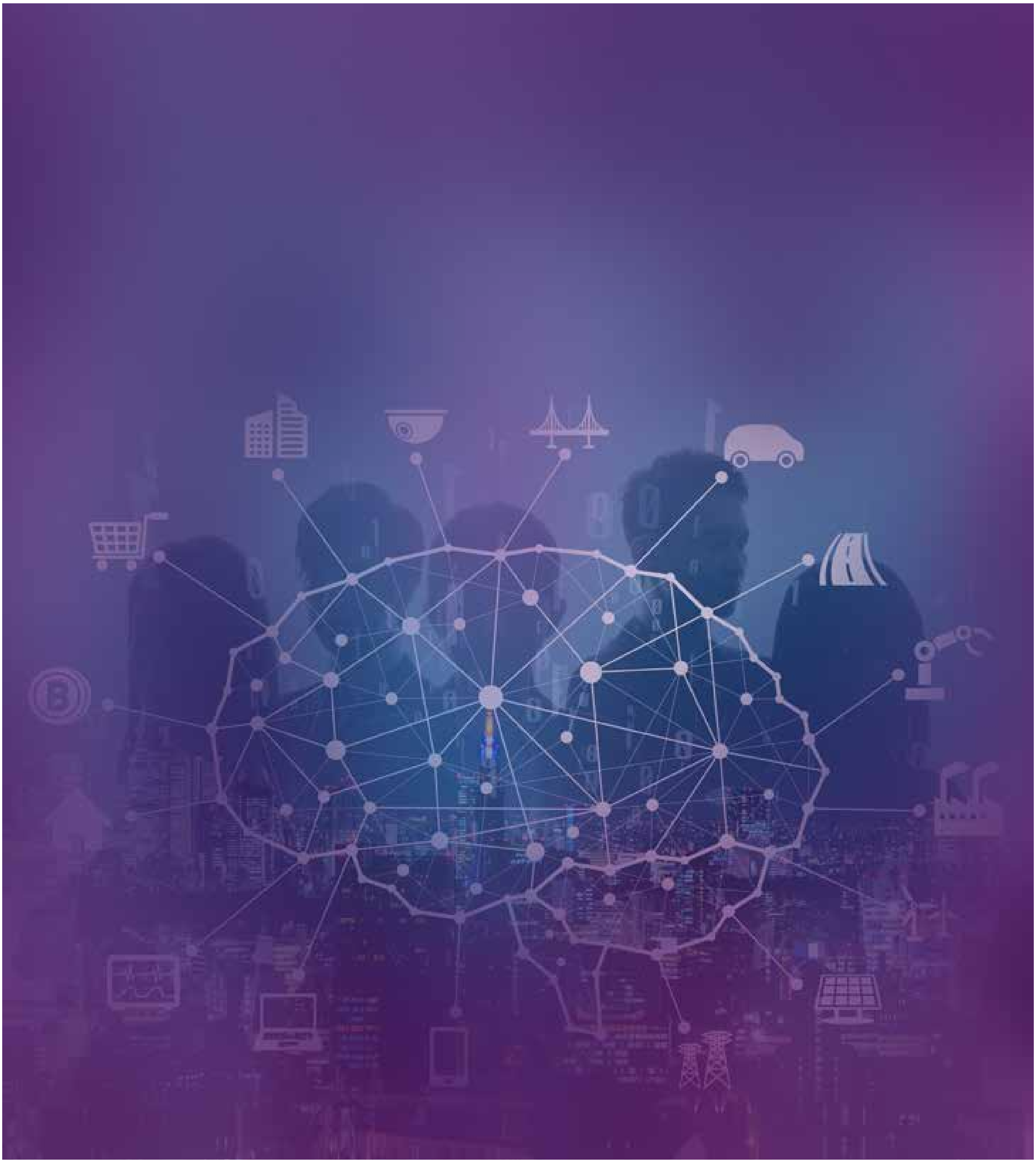
Mphasis
The Next Applied

Understanding the Credit of Technical Debt in Data Science

Whitepaper by Dr. Archisman Majumdar
AVP & Lead - Applied AI



Contents

As Machine Learning (ML) and Artificial Intelligence (AI) move into mainstream prominence, implementation of these technologies have grown exponentially. New research in ML and deep learning are enabling applications across industries and use cases. Today, ML has moved beyond proof of concept on academic, sample, and competition datasets in sandbox environments to supporting mission- and business-critical processes in diverse industries. And this has happened with great speed.

One of the challenges in speedy implementation of any new technology in production environments is technical debt. Once they hit the production servers, ML and AI systems are not immune to this.

Technical debt is a concept introduced by Ward Cunningham. Just like in a financial world, debt is necessary for speed and building of new systems. Yet, debt of the 'bad kind' can cause serious (though unintended) consequences to the financial health of individuals, firms, and economies. Using the metaphor of debt, Cunningham explained how doing things 'quick and dirty' can introduce multiple complications in the code. This, in turn, can result in a significant extra effort in later stages – leading to diminished team morale. This is part of the costs and interest we pay for incurring the debt.

To mitigate unintended potential effects of ML systems in production, researchers have introduced the concept of technical debt (as 'high interest credit cards of technical debt'), to AI and ML communities. Extensive research is under way, both in the academic and practitioner spheres. This paper aims to consolidate some key reasons for technical debt, how its potential can be identified and what are the common strategies for mitigating the challenges.

1.
Why technical debt arises in ML projects?

Technical debt arises in software development due to multiple reasons. The need for speed and deadline pressures are two important ones. Under such circumstances, teams often choose a suboptimal solution which can be implemented faster with available tools. Additionally, activities such as testing, ensuring code quality, readability, and documentation are sacrificed or compromised. This may result in system complexities and dependencies not being properly streamlined/documented.

Such scenarios are not uncommon in the world of ML and AI systems. Other reasons why ML and AI pipelines are more prone to technical debt include:

- ML models are often implemented as black boxes. This lack of explainability and inherent bias in ML models are topics of ongoing research. In production pipelines, the lack of understanding in why a particular prediction or classification is made may often lead to unintended consequences in the model or downstream applications over time
- Challenges with scalability and measurements (the art and science)
- The CACE (change anything, change everything) phenomenon
- Overfitting, feedback loops, and gradual undetected changes in the environment
- Undetected downstream consumers that may cause unpredictable behavior when the original model changes
- High data dependency, and unstable data and features
- ML has to entangle data from many applications to generate insights
- Glue code and pipeline jungles due to multiple platforms, languages and versions. ML and AI are characterized by a plethora of platforms, technologies, modeling approaches and programming languages. Very often, the language used for deployment of models is different from that used for developing them. This introduces significant risks in the system
- Stakeholder misalignment, as many data science projects start in the research labs of organizations. They may not have considered impact on all stakeholders and business areas

2.
Identifying potential ML debt in ongoing projects

When a new project is undertaken, it may be meaningful to do an analysis of the potential technical debt existing in the current implementations. This can be done by

- Understanding and closely evaluating the current data infrastructure for the issues listed in the previous section
- Checking scope boundaries and assumptions of models. Are they documented clearly? What are the levels of abstraction?
- System integration of upstream and downstream systems. When is data collected, analyzed, used?
- Monitoring, maintenance and improvement of the model on a regular basis

3.
Strategies for mitigating technical debt

In existing and new projects, some strategies for mitigating technical debt are similar to standard practices in software engineering – such as refactoring and increasing readability of the code, testing (unit, integration and regression testing), and evaluating the structure, processes and technology. For ML systems, in addition to standard techniques, technical debt may be mitigated in the following manner:

- Ongoing tests for data, feature and concept drift recognition can identify changing environments that could make earlier models obsolete or assumptions for the models invalid. Revisiting assumptions and evaluating their validity can eliminate this risk. Testing for equivalence (that is, checking for training and enabling model synchronizations) is another option. A large deviation often points to overfitting or change in the base assumption of the model features.

- Establishing a well thought out, standardized, and documented process is another mitigation technique. While some processes like CRISP-DM exist, there are also some guidelines from Google, Facebook, etc. Many organizations also use their own processes. More than the methodology, it is important to have a relevant and documented process for use by data scientists and engineers.

- Versioning and pipeline management produce derivations of datasets to ensure that multiple versions of a stage can run in parallel – while minimizing the amount of redundant computations. This can reduce the time spent in experimentation on setting up the data and keeping track of results for multiple pipelines where data is consumed. If a pipeline stage changes, then its consumers may also need to change.

- Ongoing model risk management is a technique inspired by the financial industry. This ensures ML model auditability by preserving the data that is used to build the models. Appropriate governance, policies and controls will further ensure that the veracity of data for developing models.

- It is useful to train and incentivize other departments too. Data scientists will then have an idea of the models used in production and software quality, while engineering and IT will understand the models in use (they can alert if the models are not valid anymore).
  A human-in-the-loop is a good idea even in the production environment. They can relook the predictions that the system is not completely certain about, before the data is fed into downstream applications.

4.
The path forward

While technical debt is required for growth, unpaid debt is costly. This is truer in ML and AI systems, where human intervention in the running of algorithms reduces day by day.

Building ML models is just a small part of the process of data science. Maintaining them and ensuring that they are relevant in real-world production environments is another challenge altogether. Managing the ‘process’ of data science, ML models and their deployment, and continuous evaluation of results are key to achieving the transformational potential of AI and ML solutions. Beyond hardware, software and modeling requirements, it is imperative for organizations to understand the technical debt of such projects and take adequate steps to control and optimize it.

5.
References and further reading

1. Dgrtwo/dgrtwo.github.com, Dgrtwo - *https://github.com/dgrtwo/dgrtwo.github.com/blob/master/_R/2018-05-10-scientific-debt.Rmd*
2. Hazelwood, Kim, et al. “Applied Machine Learning at Facebook: A Datacenter Infrastructure Perspective.” High Performance Computer Architecture (HPCA), 2018 IEEE International Symposium on. IEEE, 2018.
3. Mariscal, Gonzalo, Oscar Marban, and Covadonga Fernandez. “A survey of data mining and knowledge discovery process models and methodologies.” The Knowledge Engineering Review 25, no. 2 (2010): 137-166.
4. Rules Of Machine Learning: | Ml Universal Guides | Google Developers, *https://developers.google.com/machine-learning/guides/rules-of-ml/?utm source=mybridge&utm_medium=blog&utm_campaign=read_more*
5. Sculley, D., Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, and Michael Young. “Machine learning: The high interest credit card of technical debt.” (2014).
6. Techdebt 2018 International Conference on Technical Debt, *https://2018.techdebtconf.org/home*
7. The Human Cost Of Tech Debt, *https://daedtech.com/human-cost-tech-debt/*
8. van der Weide, Tom & Smirnov, Oleg & Zielinski, Michal & Papadopoulos, Dimitris & van Kasteren, Tim. (2016). Versioned Machine Learning Pipelines for Batch Experimentation. 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.
9. What Main Methodology Are You Using For Your Analytics, Data Mining, or Data Science Projects? Poll, *https://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html*

Author

Dr. Archisman Majumdar

AVP & Lead - Applied AI

Archisman is AVP & Lead - Applied AI at Mphasis NEXT Labs where he leads a cross-functional team of Data Scientists and consults Fortune 500 companies on AI and ML implementations. He holds a PhD from the Indian Institute of Management Bangalore (IIMB) in the Quantitative Methods and Information Systems area. His areas of expertise are in machine learning, product management, and information systems research.

For more information, contact: marketinginfo@mphasis.com

USA
460 Park Avenue South
Suite #1101
New York, NY 10016, USA
Tel.: +1 212 686 6655

UK
City Point, Spaces 12th Floor
1 Ropemaker Street, London
EC2Y 9HT, United Kingdom
Tel.: +44 020 7153 1327

INDIA
Bagmane World Technology Center
Marathahalli Ring Road
Doddanakundhi Village
Mahadevapura
Bangalore 560 048, India
Tel.: +91 80 3352 5000

VAS 09/07/19 US LETTER SIZE BASIL 5300

www.mphasis.com