Data Science for Blockchain: Understanding the Current Landscape

Data Science and blockchain technology are made for each other. But just how many and what kind of real-world applications are out there?

Sergey Mastitsky Apr 26 · 12 min read ★

Photo by Zdeněk Macháček on Unsplash

cryptocurrencies, and the ongoing NFT craze. From a Data Scientist's perspective, blockchains are also an exciting source of high-quality data that can be used to tackle a wide range of interesting problems using Statistics and Machine Learning. But what are those problems exactly and is there enough demand for Data Scientists in the blockchain industry to build a career? This article is an attempt to answer these questions by providing a hype-free overview of the current situation and the trends emerging at the intersection between Data Science and blockchain technology.

Disclaimer: I am neither affiliated with nor do I endorse any of the companies mentioned herein. These companies and their products are mentioned for illustrative purposes only.

First things first — terminology

There is no universally accepted definition of the term "*Data Science*". Here I will stick to my favourite definition that has been formulated by Cassie Kozyrkov (Head of Decision Intelligence at Google):

> "Data Science is the discipline of making data useful."

This is an all-encompassing definition, so it would be helpful to have a more detailed split of the ways Data Science is practised. Kozyrkov (2018) proposed the following taxonomy of Data Science applications according to the number of decisions they enable:

- *Data Mining / Descriptive Analytics* — exploratory analysis to identify interesting, previously unknown patterns in data and formulate hypotheses about the underlying process(-es) that caused those patterns. Here there are *no decisions* to be made yet — we are only "searching for inspiration" in the dataset at hand.
- *Statistical Inference* — making *one or a few* risk-controlled conclusions about the world that generalise beyond the analysed dataset.
- *Machine Learning / Artificial Intelligence* — building "decision recipes" that can be (re-)used to make *many decisions* in an automated way.

Although the difference between these types of Data Science projects is not always as clear-cut, let us accept this taxonomy as sufficient for our purposes.

According to Wikipedia,

> blocks, that are linked using cryptography.

Blockchains are implemented as decentralised, distributed, write-only databases that run on peer-to-peer computer networks. Traditionally, blockchains have been used as *ledgers* to keep records of cryptocurrency transactions (e.g., Bitcoin, Ethereum, Dash, etc.). However, nowadays this technology is gaining many other use cases.

Transactions on blockchains take place between two or more *addresses* — alphanumeric strings that serve as user pseudonyms and act similar to email addresses. One real person, the blockchain's user, can own multiple addresses. Moreover, some blockchains (e.g., Bitcoin) encourage their users to create new addresses for new transactions to maintain a high level of anonymity.

Records natively generated on a blockchain are referred to as *on-chain data*. During their analysis, such records are often enriched with *off-chain data* that can originate from any external sources (e.g., names of the entities that own certain blockchain addresses can sometimes be collected from public forums and websites).

Similar to Chen et al. (2020), we will split blockchain-related Data Science applications into two types — *"for blockchain"* and *"in blockchain"*. Applications of the first type do something useful with on-chain and possibly off-chain data, but are not necessarily deployed on the blockchain's infrastructure (e.g., an on-chain data-based dashboard that is deployed on a cloud provider's infrastructure). Applications of the second type are part of the blockchain itself.

The "in blockchain" applications can be deployed as *smart contracts*. In simple terms, a smart contract is a piece of code that resides at its own address and executes certain predefined logic in response to the contract-specific trigger(s). Smart contracts can be written in various programming languages, both general-purpose and specialised ones, such as Solidity or Vyper.

Data Science and blockchain technology are made for each other

In contrast to traditional data sources (e.g., centrally controlled corporate databases), blockchains by design provide several benefits that are important for Data Science applications:

- *High data quality*: All new records go through a rigorous, blockchain-specific validation process powered by one of the many "consensus mechanisms". Once validated and approved, these records become immutable — no one can modify

Scientist who works with such data much easier and more predictable.

- *Traceability*: Blockchain records contain all the information necessary to track their origin and context, e.g., which address initiated a transaction, time when it happened, the amount of asset transferred, and which address received that asset. Moreover, most of the public blockchains have "explorers" — websites where anyone can examine any record that has ever been generated on the respective blockchain (see, for example, the Bitcoin, Ethereum, and Ripple explorers).
- *Built-in anonymity*: Blockchains do not require their users to provide any personal information, which is important in a world where keeping one's privacy has become a real issue. From a Data Scientist's perspective, this helps to overcome the headaches associated with some of the regulations (e.g., GDPR in Europe) that require personal data to be anonymised before processing.
- *Large data volumes:* Many Machine Learning algorithms require large amounts of data to train models. This is not a problem in mature blockchains, which offer gigabytes of data.

Collecting data from blockchains is tricky, but there are options

Photo by Burst on Unsplash

individual blockchain records using the aforementioned explorer websites. However, *programmatically* collecting larger datasets suitable for Data Science purposes can be a daunting task that may require specialised skills, software, and financial resources. There are three main options one could consider.

Option 1 — Use datasets already prepared by someone else

As part of their BigQuery Public Datasets programme, Google Cloud provides full transaction histories for Bitcoin, Bitcoin Cash, Dash, Dogecoin, Ethereum, Ethereum Classic, Litecoin, and Zcash. These datasets can be easily queried using SQL and the results can be exported for further analyses and modelling. Conveniently, most of these datasets are using the same schema, making it easier to reuse SQL queries. See posts by Evgeny Medvedev here on Medium for tutorials.

There also exist static blockchain datasets that one could use for research and development purposes. Here are just a few examples:

- The Elliptic Dataset, a sub-graph of the Bitcoin graph composed of 203,769 nodes (transactions) and 234,355 edges (directed payment flows). The nodes are labelled as “licit”, “illicit” or “unknown”. This dataset has been released by the company Elliptic with the aim to spark interest in the academic and crypto communities towards building a safer cryptocurrency-based financial system (Bellei 2019; Weber et al. 2019).

- The Medalla dataset in BigQuery made publicly available by the company Nansen.ai as part of the Medalla Data Challenge run in 2020 by the Ethereum Foundation. This dataset includes variables that describe blocks and block validators on the Ethereum Beacon Chain.

- The CryptoKitties dataset, which contains attributes of thousands of “digital cats” from the famous Ethereum-based game.

- Datasets used by Reid and Harrigan (2012) and Fam and Lee (2017a, 2017b) to detect anomalous transactions on the Bitcoin blockchain.

Option 2 — Use blockchain-specific API or ETL tool

The BigQuery Public Datasets do cover major blockchain projects, but what if the blockchain of interest is not among them? The good news is that essentially all blockchains give a programmatic way to interact with their networks via the respective REST and/or Websocket APIs. See, for example, the APIs to query Bitcoin, Ethereum, EOS, NEM, NEO, Nxt, Ripple, Stellar, Tezos, TRON, Zilliqa.

or R. Examples of such libraries for Python include `bitcoin` (Bitcoin), `trinity` and `web3.py` (Ethereum), `blockcypher` (Bitcoin, Litecoin, Dogecoin, Dash), `tronpy` (TRON), `litecoin-utils` (Litecoin), etc. Examples of the R packages are fewer but do exist: `Rbitcoin` (Bitcoin), `ether` (Ethereum), `tronr` (TRON).

Introducing tronr, an R package to explore the TRON blockchain

All you need to query account balances, transactions, token transfers, and much more.

levelup.gitconnected.com

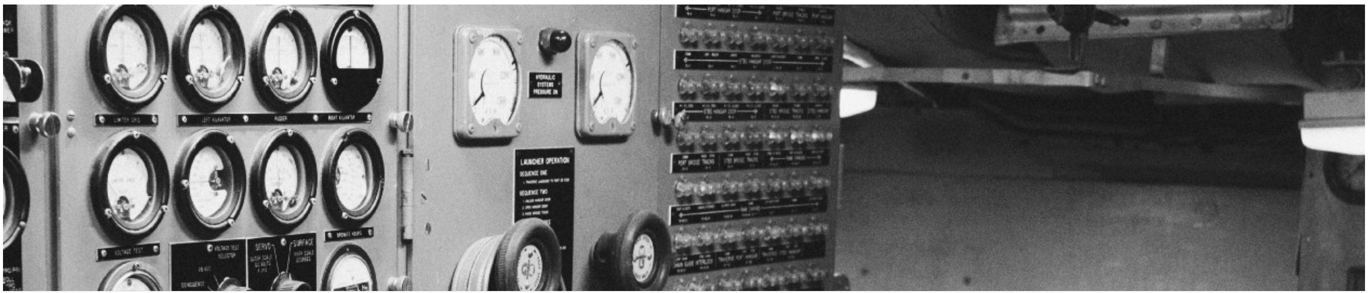
In addition to APIs, one could also consider using dedicated ETL tools to gather data from blockchains. One prominent open-source project in this space is "Blockchain ETL", a collection of Python scripts developed by Nansen.ai. In fact, these are the very scripts that feed data into the aforementioned BigQuery public datasets.

Although native blockchain APIs and open-source ETL applications give Data Scientists a lot of flexibility, using them in practice may require additional efforts and data engineering skills: setting up and maintaining a local or cloud-based blockchain node, a runtime environment to execute scripts, a database to store the retrieved data, etc. The associated infrastructural requirements may also incur substantial costs.

Option 3 — use commercial solutions

To save time, efforts, and infrastructure-related costs, one can also opt for commercial solutions for blockchain data collection. Such tools typically provide data via an API or a SQL-enabled interface using a schema that is unified across several blockchains (see, for example, Anyblock Analytics, Bitquery, BlockCypher, Coin Metrics, Crypto APIs, Dune Analytics, Flipside Crypto). This facilitates various comparative analyses and, at least in theory, makes it possible to develop Data Science applications that are interoperable across blockchains.

We live in the era of descriptive blockchain analytics



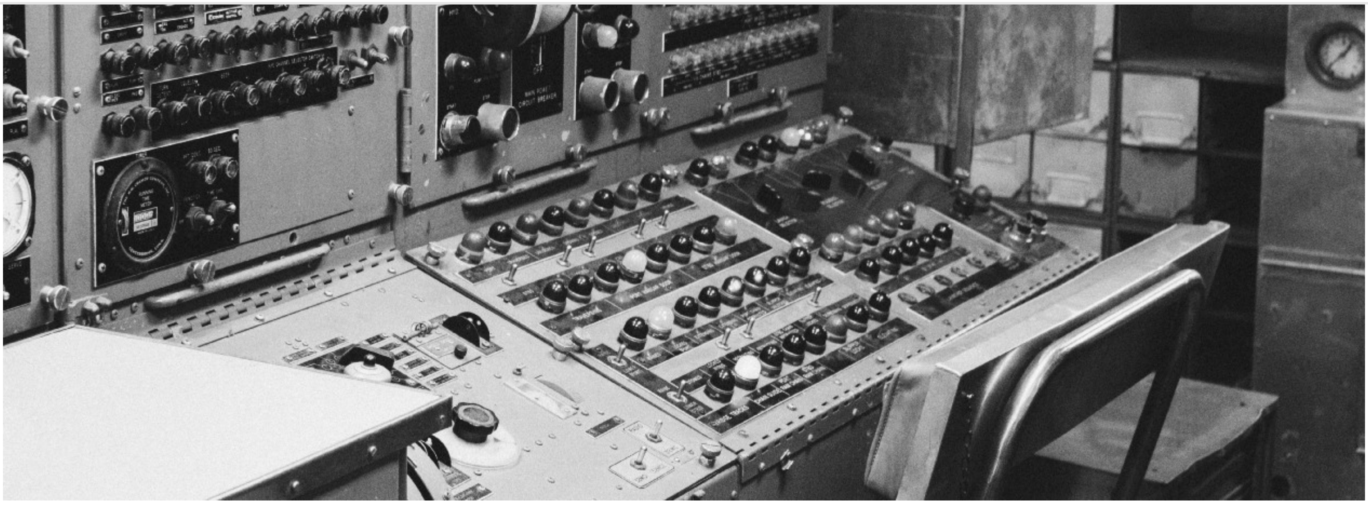


Photo by Cody Fitzgerald on Unsplash

Blockchain is still a new technology and, arguably, we only begin to understand the value of data it offers. At this early stage, it is crucial to be able to efficiently collect, summarise, and visualise data, that is perform descriptive analytics. Unsurprisingly, the majority of use cases for descriptive blockchain analytics have thus far been centred around cryptocurrencies and regulatory compliance. Numerous investigative tools have been developed to help cryptocurrency businesses, financial institutions, regulators, and law enforcement departments with the following common tasks:

- tracking and visualising the *flows of funds* between addresses in order to, for example, find stolen funds, uncover cryptocurrency price manipulations, reveal identities of thieves, and prevent money laundering;
- *labelling* the blockchain addresses and linking them to real-world entities (e.g., "darknet marketplaces", "ransomware", "scams", "mining pools", "gambling apps", "cryptocurrency exchanges", etc.);
- *real-time alerting* about on-chain events, such as suspicious transactions, large transfers of funds, activity in sanctioned addresses;
- *cross-blockchain search* using a unified data schema;
- calculating and visualising *custom on-chain metrics*;
- performing *custom data queries*.

Some of the biggest players in the investigative blockchain analytics space include AnChain.ai, Bitfury, CipherTrace, Chainalysis, Coin Metrics, Dune Analytics, Elliptic, Glassnode, Nansen.ai.

Photo by Science in HD on Unsplash

There is a large body of academic studies that applied Machine Learning to tackle various blockchain-related problems. In their comprehensive review paper, Liu et al. (2021) categorise these problems into the following three topics (see also Chen et al. 2020 for a similar discussion).

- *Cryptocurrency price prediction* — This topic is by far the most popular one, which is not surprising considering the ever-growing interest in cryptocurrencies among retail and institutional investors. The majority of published studies attempted to predict the future price of Bitcoin or Ethereum, either in absolute terms or as price direction (up or down). This has been done with algorithms ranging from simple logistic regression to XGBoost and Deep Learning-based methods (e.g., Greaves and Au 2015; Abay et al. 2019; Barry and Crane 2019; Chen et al. 2020). Model inputs typically included a combination of on-chain (e.g., number of active addresses, transaction volume, mining difficulty, transaction graph metrics) and off-chain variables (e.g., trading volumes on exchanges). Overall, neural networks (in particular, those with LSTM-based architectures) have been found to outperform tree-based algorithms, although the accuracy of prediction even in the best-performing models was only marginally higher than a random guess.

predefined groups or, better yet, link them to real-world entities can be of great value. This is especially true for addresses involved in illicit activities, such as money laundering, distribution of drugs, ransomware, Ponzi schemes, human trafficking, and even terrorism financing. The problem of address categorisation has been successfully tackled with supervised learning in several studies that developed either binary classifiers (e.g., “illicit vs licit” as in the aforementioned paper by Weber et al. 2019) or multiclass classifiers (e.g., “exchange”, “pool”, “gambling”, and “service” addresses as in Liang et al. 2019; see also Harlev et al. 2018, Yin et al. 2019; Michalski et al. 2020). The respective models have been developed using on-chain data and a variety of standard Machine Learning algorithms. Interestingly, in contrast to the cryptocurrency price forecasting studies, tree-based methods (in particular, Random Forest) have often outperformed Deep Learning-based algorithms (Liu et al. 2021).

- *Anomaly detection*— Blockchains can be subject to many kinds of malicious attacks and fraudulent behaviours (e.g., Moubarak et al. 2018; Rahouti et al. 2018), which potentially can be detected by analysing the transaction patterns. As the number of abnormal transactions is naturally small, this problem has been addressed with either empirically derived rules or unsupervised Machine Learning methods, such as *k*-means clustering, one-class SVM, Mahalanobis distance-based labelling, and Isolation Forest (e.g., Camino et al. 2017; Pham and Lee 2017a, 2017b). Liu et al. (2021) conclude that most of these studies suffered from a low recall, thus warranting the need for further research. At the same time, Tann et al. (2019) have built a highly accurate LSTM-based binary classifier to detect code vulnerabilities in smart contracts.

In addition to the use cases described above, many researchers have proposed and even experimentally tested Machine Learning-powered blockchain platforms aimed at improving the existing electronic health record management systems, enabling better traceability in supply chains, increasing security in the Internet of Things networks, etc. (Salah et al. 2018; Chen et al. 2020). Several papers have also proposed blockchain-based platforms and protocols for collective training and dissemination of Machine Learning models (e.g., “DInEMMo” by Marathe et al. 2018, “DanKu protocol” by Kurtumlus and Daniel 2018, and “Decentralized and Collaborative AI on Blockchain” by Harris and Waggoner 2019).

Despite the numerous examples of experimental Machine Learning-powered blockchain systems seen in academic literature, practical implementations of such systems are still rare. There are many reasons for this, both data-related and infrastructural (Salah et al.

labels are usually available only for a small proportion of addresses, which impedes the use of supervised learning (but see Rodriguez 2019 for a discussion of possible solutions).

Many Machine Learning algorithms can potentially be deployed in blockchains via smart contracts thanks to the commonly used plain-text formats PMML, PFA, and ONNX (Wang 2018). However, deploying some of the more sophisticated algorithms, such as a Tensor Flow-based deep neural network trained on GPUs, is still a non-trivial task.

It can also be difficult to deploy and use Machine Learning models in blockchains due to the low bandwidth and high cost of transactions, lack of technical standards and interoperability protocols, and the need for trusted suppliers of external data (a.k.a. "oracles") that can make Machine Learning-enabled smart contracts more useful (Salah et al. 2018).

Nevertheless, a number of companies, including Algorithmia, AnChain.ai, Bitfury, Fetch.ai, IBM, IntoTheBlock, SingularityNET and others, claim that they are using Machine Learning in their blockchain products. No doubt, we will see more of such companies and products in the near future. Among other developments, this growth will be greatly facilitated by easier ways of making smart contracts "smarter" with off-chain signals provided by oracles — Chainlink and Provable are already offering the respective solutions.

Conclusion: if you want to become a blockchain Data Scientist, now is the perfect time

Blockchain technology has the potential to transform many industries and business processes. In their recent article, the Forbes Technology Council have identified 13 evolving and emerging use cases for blockchain, including rights management for artists, cross-industry data consolidation, decentralised finance, supply chain management, user authentication and password management, electronic health records, etc. All of these developments will require an army of experts who are capable of "making data useful", that is Data Scientists. The range of interesting and unsolved blockchain Data Science problems is enormous. Moreover, many of those problems are yet to be formulated. Thus, if you are thinking about entering the exciting world of blockchain as a Data Scientist, the timing could not be better. Many of the companies mentioned in this article already have open positions for Data Scientists— do check out the "Careers" sections on their websites!

I provide Data Science consulting services. Get in touch!

Blockchain Data Science Machine Learning Predictive Analytics Artificial Intelligence