Property-Based Testing of Data Analysis Scripts
A Focus on Hypothesis for Python

Jean-Sebastian de Wet, Jan-Philipp Kiel, and Pascal Mager

University of Cologne, Cologne, Germany

Abstract. This paper explores property-based testing as a method to ensure data analysis scripts' reliability, especially in DLR research using Python, Pandas, and Matplotlib. It outlines challenges with traditional testing in scenarios with diverse data values, emphasizing the need for innovative testing strategies. The paper thoroughly covers property-based testing, including its history, key principles, use cases, and integration in the test pyramid. It then focuses on Hypothesis for Python, a powerful tool for property-based testing, discussing its use, integration with pytest, and unique features. Real-world application is demonstrated with code examples, highlighting how property-based testing, especially with Hypothesis, strengthens data analysis scripts' reliability. The paper concludes by summarizing key findings and emphasizing the crucial role property-based testing, like Hypothesis, plays in boosting researchers' confidence with unknown data.

Keywords: Property-Based Testing · Data Analysis Scripts · Hypothesis · Python · pytest · Reliability · Test Pyramid · Code Examples · DLR Research.

1 Introduction

In the evolving landscape of data analysis, the complexity and volume of datasets have grown exponentially [27], presenting unique challenges across various fields. This surge in data complexity necessitates robust testing methodologies to ensure the accuracy and reliability of data analysis tools and scripts. Traditional testing approaches, primarily based on specific input-output cases, often fall short in addressing the dynamic and unpredictable nature of modern datasets. These limitations are particularly evident in specialized fields like aerospace research, where the data's scope and diversity are exceptionally vast.

2 Background

Within the German Aerospace Center (DLR), the reliability of data analysis scripts is a cornerstone of successful research outcomes. DLR researchers frequently use Python [17], along with its powerful libraries like Pandas[1] and Mat-

[1] https://pandas.pydata.org/, accessed: 21.01.2024

plotlib[2], for complex data manipulations and visualizations. This introduces significant testing challenges. DLR, being at the forefront of aerospace research and development, deals with an enormous range of data variables, from satellite imagery to flight dynamics.[3] This variety and complexity of data make testing particularly challenging. This paper explores the adoption of property-based testing, presenting it as an innovative and essential strategy to overcome these testing challenges, especially in scenarios involving large possible value ranges of data.

3 Method

4 Results

4.1 Overview of Property-Based Testing

History of Property-Based Testing The origins of property-based testing (PBT) can be traced back more than 20 years ago, even before 2000. Although it had already been a topic within information technology research like in Goldreich (1998) and Fink (1997), it gained much more attention through the development of QuickCheck[4] [26, 9, 21, 12]. Beginning with research questions concerning topics like automation of test input generation and automated techniques in general [9] and guiding the automated input generator towards values with higher probability of failure [19], more recent papers deal with the implementation of different frameworks or platforms and techniques of PBT-application [22, 12, 26, 7]. To top it off, PBT already enjoys wide-ranging support in different programming languages including automation capabilities [5, 22, 12, 8, 26], as well as the application within many different python projects [7].

How Property-Based Testing Works PBT is a method enabling the formal verification of a software [5, 9, 12, 24], with the key concept of validating high-level or general properties of a software [9, 19, 12, 7]. Test cases used within the application of this method, are usually formulated using logical descriptions of a software's expected behaviour [5, 9, 12, 19, 7]. More explicitly, these tests may include pre- or post-conditions of the system [12]. In order to provide formal validation of the system's behaviour, a single test case is executed many times with randomly generated input in search for counterexamples resulting in violation of a specific property or even a crash of a software, therefore invalidating said property [5, 19, 22, 8, 24, 7]. For randomly generating input, data generators are used [5, 19, 22, 8] which can be adjusted according to "domain-specific knowledge" [5]. Through this automated execution of tests with random input PBT tries to approximate the validity of a certain property, as it has to withstand the

[2] https://matplotlib.org/, accessed: 21.01.2024
[3] https://www.dlr.de/en/dlr/about-us, accessed: 21.01.2024
[4] https://www.cse.chalmers.se/ rjmh/QuickCheck/, accessed: 21.01.2024

check using many different instantiations within a given input range or otherwise the property is falsified [9, 8, 7, 24]. To further elaborate on the properties, which represent the desired behaviour in terms of input / output of given tested functions through specifications [5, 9, 19], some examples can be given. To start with a simple one, think of a function that adds two numbers (A, B) and returns the sum of both numbers (A + B). Hier ggfs. Beispiel für sortierter Liste einfügen –> Beinhaltet "Invariante" You can define a test which asserts that for any given input for either A and B, the function will return the addition of both numbers [7]. Another example which is frequently used for showcasing PBT are binary search trees [7, 26]. In PBT "one desirable property [could be], that if we [you] insert a key into a valid BST, then it should remain a valid BST" [26]. In the context of PBT you would then use logic expressions for your tests and use a random input generator to check whether the given property is violated. Other possible use cases could be software-security related as example in the case of authentication [9] or "the correctness of hardware [and] the external software involved" [5]. Long story short, PBT allows for the formal verification of a system's invariants [9, 8, 7].

Advantages and Disadvantages of Property-Based Testing As already mentioned, PBT allows to formally validate the correctness of a software by testing specified properties using randomly generated input for each test. However, describing all of a system's expected behaviour in a logical style is paired with reduced feasibility [5, 16]. By applying PBT, required endeavour for formal validation can be lowered [13, 5, 24]. Moreover, specifications used for PBT might also improve cooperation between software engineers (SE) and software testers in larger projects, as the language used for defining tests is easier to grasp compared to abstract proofs [5, 19]. Besides this, PBT can also be applied to test in "multiple domains" [15] - to name a few examples next to the given examples of the previous chapter: interfaces [15, 10, 18], e.g. by testing invariants regarding the responses of requested URLs of REST-APIs [15]. Other possible domains are telecom systems [3], file synchronisation services [14] and databases [4]. Despite its wide applicability and advantages regarding formal validation, it lowers engineering effort in terms of defining individual test cases as well as input parameters [5, 19, 7] and might as well offer incentives for SEs to design the code to be easily expressed by properties [5]. More specifically, when compared to manually written tests like in unit-testing, PBT allows the SE to put more emphasis on ensuring and restoring correctness of the software and less on "defining test case inputs, examples, and scenarios" [7], which is also less "mundane" [19]. It therefore reduces costs related to testing including change induced costs [5, 19]. Furthermore it allows for validating a software based on a much larger range of inputs [19] [Hypothesis for Software Testing Research] and even more creative or sophisticated inputs [4]. Therefore PBT complements traditional testing techniques by unveiling yet unknown bugs within even well tested systems [4, 14, 3] and in general, is useful for finding bugs within the implementation of a software and its specifications [5, 9, 19, 24, 6, 7].

Although PBT offers quite some advantages, it does not come without any disadvantages. Due to the randomness of the input generator provided by tools the chances of finding more specific bugs reduces depending on the portion of erroneous inputs of the entire input range and therefore might fail to unveil errors [19, 22, 8, 26]. Not to mention that an enormous amount of tests processed implies reduced efficiency [8, 26], as you try to approximate a formal proof using many randomly picked scenarios [9, 8, 24]. Löscher (2017) gave quite a good fictional example using a "system of network nodes" [19]. They tried to falsify the property that for any input scenarios (graphs created), the longest of the shortest paths "between the sink and other nodes [...] should not exceed 21 hops" [19]. Even after "100000 tests" [19] they were not able to falsify the property, which could be done "by hand" [19]. A possible solution to this problem is implied by the usage of individually conceptualised data generators [19, 8, 26, 24, 6] or targeted property-based testing [19]. While constraining data generators by using domain knowledge in order to reduce the e.g. by using pre-conditions to cancel a test [19, 8, 26], targeted property-based testing tries to guide the input generator "with search techniques towards values, that have a higher probability of falsifying a property" [19]. Especially the first mentioned technique comes with its own challenges caused by the required development of your own generators, resulting in a reduced attractiveness of PBT [19, 8, 26].

Position in the Test Pyramid In order to position PBT at a level of the testing pyramid, we focus on the level of unit and integration (service) testing [1, 25]. To the best of our knowledge, PBT has not been applied to test the highest lvl of either system or UI tests [25, 1], thus the highest lvl of the testing pyramid is ignored in discussing the position of PBT within the test pyramid. The relevant levels are now shortly summed up.

Beginning with unit testing, "developers perform unit testing to ensure that each component correctly implements its design and is ready to be integrated into a system of components" [integration testing]. In other words, "testing is performed in isolation from other components" [integration testing] and focuses on "example-based" [7] testing of individual parts – the "smallest parts" [1] – of a system [11, 7].

Within integration tests, every component of a system is integrated with each other if needed, including relevant external components [1, 11, 25]. The goal is to ensure that interaction between the given components works correctly and are therefore properly integrated [11, 1].

Although most of the mentioned use cases of chapter 2 (how does PBT work) relate to individual components in the means of functions such as adding numbers, inserting nodes in binary search trees and the sorted list [müssen hier die indirekten Zitate rein?], PBT also offers possible application surrounding the physical and external components of a software [5]. Coming back to the authentication-functionality provided by a system, you can not only test the functionality of authentication but also the integration with said authentication services [9]. Furthermore PBT has already been applied in many different cases

of testing RESTful APIs in the context of OpenAPIs [15], telecom systems [3], synchronisation services [14] or AUTOSAR software [4]. In the case of QuickREST you can use the response codes in order to differentiate between invalid and URL-requests, which in the end lead to revealing a yet unknown "underspecification" [15] of a used OpenAPIS documentation [15]. Whereas in the case of AUTOSAR the testing of correct request processing unveiled a yet unknown bug in terms of task prioritisation [4]. Last but not least, entire crashes of addressed components can be perceived as well [3].

Therefore the method of PBT can be located within the level of unit and integration testing. It can be very well applied to testing individual components of a system, but is not limited to it, because testing the integration of components is also possible and is used for both in practice.

Tools and Programming Languages As already mentioned, PBT enjoys wide ranging support in many different programming languages [5, 26]. It is supported by Java (Quicktheories[5]), coq (QuickChick[6]), Scala (ScalaCheck[7]), Erlang (QuickCheck[8] and PropEr[9]), Haskell (QuickCheck[10]), OCaml (QCheck[11] and Crowbar[12]) to name a few examples [20, 22, 24, 2, 23, 6]. Obviously many of these tools were inspired by QuickCheck, being the tool popularising PBT. However in this work we focus on Hypothesis[13], a PBT implementing framework for Python. It is a framework receiving much attention recently [7, 21], which is compatible with py.test, unittest and "probably many others", while being open source "under the Mozilla Public License 2.0".[14]

4.2 Introduction to Hypothesis for Python

Overview of Hypothesis

- Provide a brief introduction to Hypothesis for Python.
- Mention its key features and advantages.

How to Use Hypothesis

- Write a mini how-to guide on using Hypothesis, including integration with pytest.
- Include code snippets for better understanding.

[5] https://github.com/quicktheories/QuickTheories, accessed: 21.01.2024
[6] https://github.com/QuickChick/QuickChick, accessed: 21.01.2024
[7] https://scalacheck.org/, accessed: 21.01.2024
[8] http://www.quviq.com/products/erlang-quickcheck/, accessed: 21.01.2024
[9] https://proper-testing.github.io/, accessed: 21.01.2024
[10] https://hackage.haskell.org/package/QuickCheck, accessed: 21.01.2024
[11] https://github.com/c-cube/qcheck/, accessed: 21.01.2024
[12] https://github.com/stedolan/crowbar, accessed: 21.01.2024
[13] https://hypothesis.works/, accessed: 21.01.2024
[14] https://hypothesis.works/products/, accessed: 21.01.2024

4.3 Main Concepts and Features of Hypothesis

Strategies and Data Generation

- Explain the concept of strategies in Hypothesis for generating test data.

Property-Based Testing with pytest

- Detail how Hypothesis integrates with pytest.
- Provide examples of test functions using Hypothesis.

Data Analysis Applications and Benefits

- Discuss how data analysis applications can benefit from Hypothesis.
- Reference specific features, such as the support for NumPy.

4.4 Application of Hypothesis in Data Analysis

Code Examples

- Provide practical code examples demonstrating the use of Hypothesis in data analysis scripts.
- Showcase scenarios where property-based testing adds value.

```
import numpy as np

def incmatrix(genl1,genl2):
    m = len(genl1)
    n = len(genl2)
    M = None #to become the incidence matrix
    VT = np.zeros((n*m,1), int)  #dummy variable

    #compute the bitwise xor matrix
    M1 = bitxormatrix(genl1)
    M2 = np.triu(bitxormatrix(genl2),1)

    for i in range(m-1):
        for j in range(i+1, m):
            [r,c] = np.where(M2 == M1[i,j])
            for k in range(len(r)):
                VT[(i)*n + r[k]] = 1;
                VT[(i)*n + c[k]] = 1;
                VT[(j)*n + r[k]] = 1;
                VT[(j)*n + c[k]] = 1;

                if M is None:
```

```
            M = np.copy(VT)
        else:
            M = np.concatenate((M, VT), 1)

        VT = np.zeros((n*m,1), int)

    return M
```

Illustrative Cases

- Present specific cases where Hypothesis helped discover issues in data analysis scripts.

5 Discussion

6 Conclusion

- Summarize the key points discussed in the paper.
- Emphasize the importance of property-based testing, particularly with tools like Hypothesis, in enhancing the reliability of data analysis scripts.

References

1. Aniche, M.: Effective software testing. Manning Publications Co, Shelter Island, NY, 1 edn. (2022), includes bibliographical references and index
2. Arts, T., Castro, L.M., Hughes, J.: Testing erlang data types with quviq quickcheck. In: Proceedings of the 7th ACM SIGPLAN workshop on ERLANG. ICFP08, ACM (Sep 2008). https://doi.org/10.1145/1411273.1411275
3. Arts, T., Hughes, J., Johansson, J., Wiger, U.: Testing telecoms software with quviq quickcheck. In: Proceedings of the 2006 ACM SIGPLAN workshop on Erlang. ICFP06, ACM (Sep 2006). https://doi.org/10.1145/1159789.1159792
4. Arts, T., Hughes, J., Norell, U., Svensson, H.: Testing autosar software with quickcheck. In: 2015 IEEE Eighth International Conference on Software Testing, Verification and Validation Workshops (ICSTW). IEEE (Apr 2015). https://doi.org/10.1109/icstw.2015.7107466
5. Chen, Z., Rizkallah, C., O'Connor, L., Susarla, P., Klein, G., Heiser, G., Keller, G.: Property-based testing: Climbing the stairway to verification. In: Proceedings of the 15th ACM SIGPLAN International Conference on Software Language Engineering. SLE '22, ACM (Nov 2022). https://doi.org/10.1145/3567512.3567520
6. Claessen, K., Hughes, J.: Quickcheck: a lightweight tool for random testing of haskell programs. In: Proceedings of the fifth ACM SIGPLAN international conference on Functional programming. ICFP00, ACM (Sep 2000). https://doi.org/10.1145/351240.351266
7. Corgozinho, A.L., Valente, M.T., Rocha, H.: How developers implement property-based tests. In: 2023 IEEE International Conference on Software Maintenance and Evolution (ICSME). IEEE (Oct 2023). https://doi.org/10.1109/icsme58846.2023.00049

8. Elazar Mittelman, S., Resnick, A., Perez, I., Goodloe, A.E., Lampropoulos, L.: Don't go down the rabbit hole: Reprioritizing enumeration for property-based testing. In: Proceedings of the 16th ACM SIGPLAN International Haskell Symposium. Haskell '23, ACM (Aug 2023). https://doi.org/10.1145/3609026.3609730
9. Fink, G., Bishop, M.: Property-based testing: a new approach to testing for assurance. ACM SIGSOFT Software Engineering Notes **22**(4), 74–80 (Jul 1997). https://doi.org/10.1145/263244.263267
10. Francisco, M.A., López, M., Ferreiro, H., Castro, L.M.: Turning web services descriptions into quickcheck models for automatic testing. In: Proceedings of the twelfth ACM SIGPLAN workshop on Erlang. ICFP'13, ACM (Sep 2013). https://doi.org/10.1145/2505305.2505306
11. Hartmann, J., Imoberdorf, C., Meisinger, M.: Uml-based integration testing. In: Proceedings of the 2000 ACM SIGSOFT international symposium on Software testing and analysis. ISSTA00, ACM (Aug 2000). https://doi.org/10.1145/347324.348872
12. Honarvar, S., Mousavi, M.R., Nagarajan, R.: Property-based testing of quantum programs in q#. In: Proceedings of the IEEE/ACM 42nd International Conference on Software Engineering Workshops. ICSE '20, ACM (Jun 2020). https://doi.org/10.1145/3387940.3391459
13. Hritcu, C., Lampropoulos, L., Spector-Zabusky, A., de Amorim, A.A., Dénès, M., Hughes, J., Pierce, B.C., Vytiniotis, D.: Testing noninterference quickly. Journal of Functional Programming **26** (2016). https://doi.org/10.1017/s0956796816000058
14. Hughes, J., Pierce, B.C., Arts, T., Norell, U.: Mysteries of dropbox: Property-based testing of a distributed synchronization service. In: 2016 IEEE International Conference on Software Testing, Verification and Validation (ICST). IEEE (Apr 2016). https://doi.org/10.1109/icst.2016.37
15. Karlsson, S., Causevic, A., Sundmark, D.: Quickrest: Property-based test generation of openapi-described restful apis (2019). https://doi.org/10.48550/ARXIV.1912.09686
16. Koopman, P., Achten, P., Plasmeijer, R.: Model Based Testing with Logical Properties versus State Machines, pp. 116–133. Springer Berlin Heidelberg (2012). https://doi.org/10.1007/978-3-642-34407-7_8
17. von Kurnatowski, L., Schlauch, T., Haupt, C.: Software development at the german aerospace center: Role and status in practice. In: Proceedings of the IEEE/ACM 42nd International Conference on Software Engineering Workshops. ICSE '20, ACM (Jun 2020). https://doi.org/10.1145/3387940.3392244
18. Lamela Seijas, P., Li, H., Thompson, S.: Towards property-based testing of restful web services. In: Proceedings of the twelfth ACM SIGPLAN workshop on Erlang. ICFP'13, ACM (Sep 2013). https://doi.org/10.1145/2505305.2505317
19. Löscher, A., Sagonas, K.: Targeted property-based testing. In: Proceedings of the 26th ACM SIGSOFT International Symposium on Software Testing and Analysis. ISSTA '17, ACM (Jul 2017). https://doi.org/10.1145/3092703.3092711
20. MacIver, D.: Quickcheck in every language (Apr 2016), https://hypothesis.works/articles/quickcheck-in-every-language/
21. MacIver, D., Hatfield-Dodds, Z., Contributors, M.: Hypothesis: A new approach to property-based testing. Journal of Open Source Software **4**(43), 1891 (Nov 2019). https://doi.org/10.21105/joss.01891
22. Padhye, R., Lemieux, C., Sen, K.: Jqf: coverage-guided property-based testing in java. In: Proceedings of the 28th ACM SIGSOFT International Symposium on Software Testing and Analysis. ISSTA '19, ACM (Jul 2019). https://doi.org/10.1145/3293882.3339002

23. Papadakis, M., Sagonas, K.: A proper integration of types and function specifications with property-based testing. In: Proceedings of the 10th ACM SIGPLAN workshop on Erlang. ICFP '11, ACM (Sep 2011). https://doi.org/10.1145/2034654.2034663
24. Paraskevopoulou, Z., Hriţcu, C., Dénès, M., Lampropoulos, L., Pierce, B.C.: Foundational Property-Based Testing, pp. 325–343. Springer International Publishing (2015). https://doi.org/10.1007/978-3-319-22102-1_22
25. Radziwill, N., Freeman, G.: Reframing the test pyramid for digitally transformed organizations (2020). https://doi.org/10.48550/ARXIV.2011.00655
26. Shi, J., Keles, A., Goldstein, H., Pierce, B.C., Lampropoulos, L.: Etna: An evaluation platform for property-based testing (experience report). Proceedings of the ACM on Programming Languages **7**(ICFP), 878–894 (Aug 2023). https://doi.org/10.1145/3607860
27. Taylor, P.: Amount of data created, consumed, and stored 2010-2020, with forecasts to 2025 [infographic]. Statista (Nov 2023), https://www.statista.com/statistics/871513/worldwide-data-created/