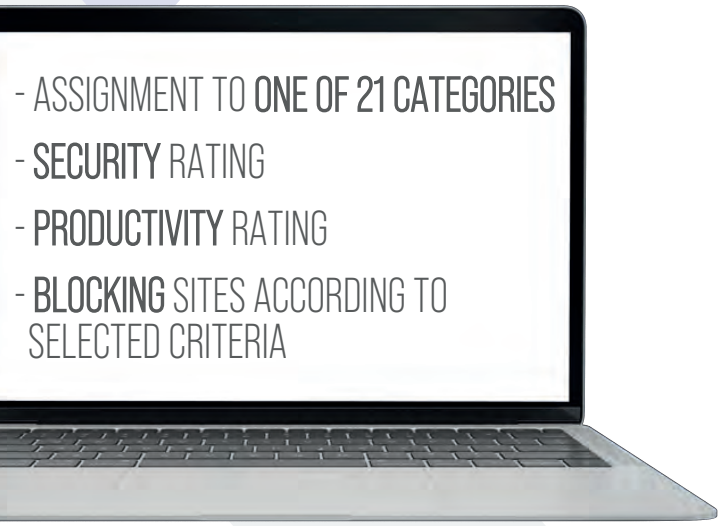
BTC MACHINE LEARNING **WEB CLASSIFICATION**

DATASHEET

eAuditor V7 WEB



21 CATEGORIES

Pages are classified based on their actual content in a repeatable manner and are assigned to one of 21 categories.

EFFECTIVENESS > 95%

The use of artificial intelligence algorithms and a built-in database of keyword patterns is responsible for high effectiveness.

API

An API is available for external entities.

24H/DAY, 365 DAYS A YEAR

The classifier works all the time in real time.

MACHINE LEARNING IN WEBSITE CLASSIFICATION

Website classification can be useful in any entity where supervision and control of user activity can have **a real impact on security**. Implementation of the machine learning algorithm **allows for efficient and fast classification of each website** in terms of its content and assignment (classification) to the appropriate category. The website classification module in the eAuditor V7 WEB system is prepared for the occurrence of various random events in such a way that despite a server error or the expiration of the website, it does not interrupt its operation and correctly performs its task, **assigning websites to the appropriate categories.**

WEBSITE CLASSIFIER

The Bayesian classifier, which is based on Bayes' theorem, is particularly suitable for solving problems with very high dimensions at the input. Despite the simplicity of the method, it often works better than other, very complicated classification methods. The aforementioned classifier can be trained in supervised learning mode. This means that for the algorithm to work correctly and even better, human supervision is necessary, who will constantly analyze and correct any errors in the algorithm. **The classification is correct as long as the correct category is more probable than others. Support for the algorithm's operation is provided by BTC.**

In practice, it happens that the algorithm will indicate a different category than we expect. This happens especially on news pages, which consist of many articles on many topics and industries. Then the algorithm may indicate the wrong category.

WEBSITE RATING TIME

Classification of a single URL takes from **1 to 2 seconds**. In practice, higher efficiency is achieved due to multi-threaded handling of classification processes (simultaneous classification of several dozen or several hundred pages).

CORRECTNESS OF WEBSITE CLASSIFICATION

As part of the machine learning test in eAuditor V7 WEB, **1000 random and unpopular websites were categorized. The correctness** of assigning categories to these websites is **about 90%**. The problem with achieving better results does not lie with the algorithm, because it determines the highest probability of a given category. The problem is that one website can be included in several categories at once and each of the categories can be correct.

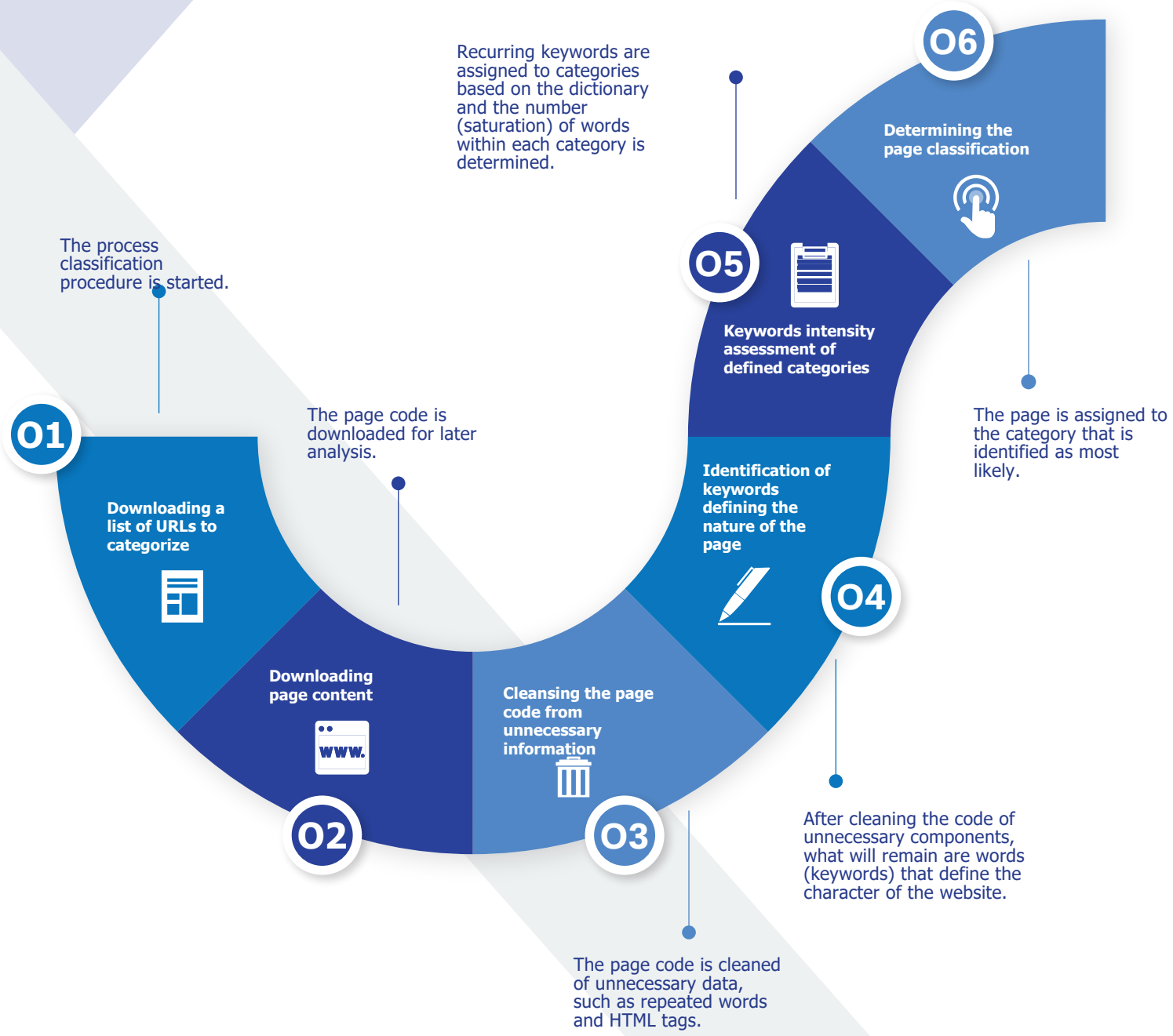
For example, *www.onet.pl* can be categorized as both news and media, as well as entertainment or law and politics.

EFFICIENCY

The classifier engine is hosted in the data center and is supported by a scalable set of servers, which ensures **unlimited performance.**

BTC MACHINE LEARNING **WEB CLASSIFICATION**

HOW DOES MACHINE LEARNING WORK IN WEBSITE CLASSIFICATION?



WHY DID WE INTRODUCE MACHINE LEARNING TO EAUDITOR V7 WEB?

Here are some reasons why we used Machine Learning in eAuditor V7 WEB instead of a web classification database:

- the database of websites with assigned categories takes up a lot of space (over 1 TB),
- the number of websites is not several thousand or even millions, currently it is a number difficult to estimate,
- websites can change their category faster than ready databases of website categories,
- databases require constant updating, which is expensive and takes a lot of time,
- machine learning categorizes websites individually, according to the needs of each user,
- the classifier copes perfectly with the specifics of the Polish language.