

# LEARNING USING PRIVILEGED INFORMATION (LUPI)

## A Revolutionary Paradigm for Data Analytics and Machine Learning

### WHY DATA ANALYTICS AND MACHINE LEARNING ARE HARD

The era of “big data” has given rise to a huge spectrum of opportunities for making businesses more efficient and responsive to their customers’ needs by predicting outcomes and trends. Real-time or longer-term insights can be obtained by analyzing the myriads of information available from open sources as well as proprietary databases. Machine learning tools have flooded the marketplace for analyzing and making sense of all this data.

Despite the rapid advances and many successes in machine learning for analytics, significant challenges remain. Available data is often incomplete and inaccurate and data about specific outcomes of particular interest may be limited in volume or expensive to acquire. Business owners must be able to not only obtain data analysis results, but also gauge the accuracy of the results and derive actionable intelligence from these results. To meet the needed levels of accuracy, machine learning systems may require lengthy and computationally-intensive training, which is time-consuming and expensive to perform.

### THE CHALLENGE: BIG DATA ISN'T ALWAYS BIG ENOUGH

Machine learning and other data analytics algorithms typically work by computing statistical properties of the data, and using these statistical properties to make decisions, including making predictions, characterizing markets and customers, tracking trends, and so on. As we all know, the science of statistics is predicated on the availability of big data: the more data you have, the more accurately you can characterize the statistical properties of the data. The availability of cheap computing power and storage has made big data and its analysis a reality.

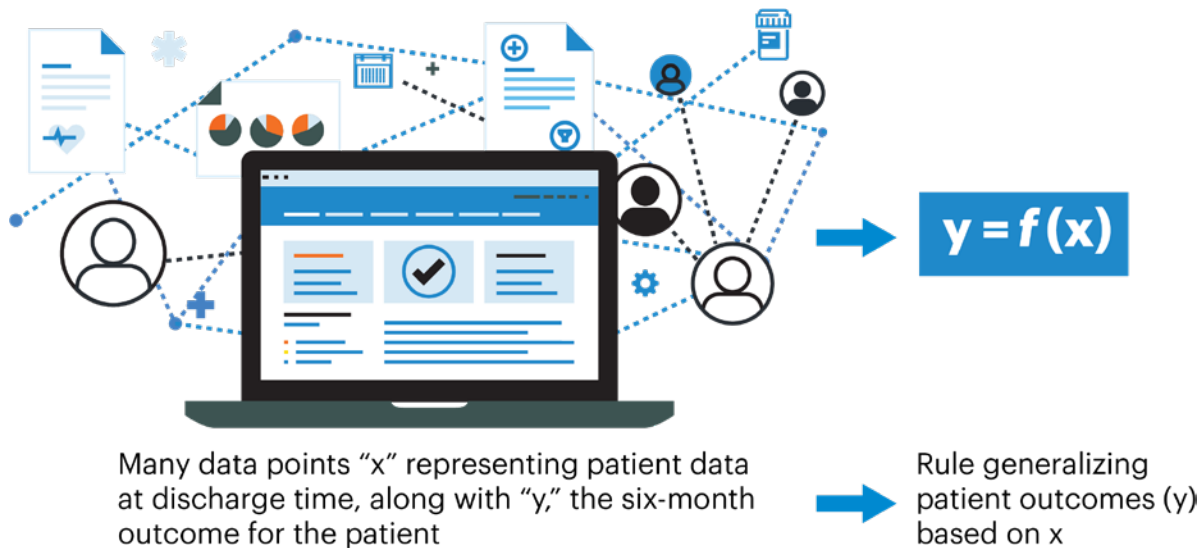
As an example, a hospital may want to analyze its patient records to try and predict the probability that a discharged patient will be re-admitted to the hospital within the next six months.

Analytics algorithms typically use historical data about patients to generate a statistical model of patient outcomes, which can then be used for predictions. This model generation process is one of learning, as the algorithm learns about patient outcomes by analyzing historical data about past outcomes. This historical data contains patient data that can be divided into two categories: category 1 contains patients who were re-admitted to the hospital within six months of their initial stay, and category 2 contains patients who weren't. However, the number of patients in the first category is typically several orders of magnitude smaller than the number of patients in the second category. Thus, the amount of data about patients in the first category is severely limited. Since most analytics algorithms require large amounts of data to be able to draw statistically valid conclusions, this makes it difficult to obtain a high level of accuracy in predicting patient re-admittance.

### THE LUPI SOLUTION: SOMETIMES LESS IS MORE

So, how do we address this issue? In a nutshell, Peraton Labs has an advanced analytics approach—based on the Learning Using Privileged Information (LUPI) paradigm—that mimics the process of human learning. By taking advantage of so-called privileged information, LUPI learns both more quickly and more accurately, requiring many fewer training examples.

Note first, in the hospital example, that a doctor who analyzes a complete patient record can typically predict the category to which this patient belongs (category 1 or 2) much more accurately than existing analytics algorithms. Unfortunately, doctors simply don't have the necessary time to repeatedly review all of the patient data, perform this kind of comprehensive analysis, and update the patient prediction again and again. LUPI works to capture the knowledge of a doctor and incorporate it into an analytics algorithm that provides accurate predictions.



To understand the LUPI solution at an intuitive level, we need to understand how current analytics algorithms analyze data. For the hospital example, the data provided to analytics algorithms is a collection of patient records. Each patient record contains data about the patient, and an indication of whether the patient belongs to category 1 (re-admitted within 6 months) or category 2 (not re-admitted within 6 months). The figure above illustrates the analysis process.

A technical point to note about the process illustrated above is the following: the analytics algorithms learn a model that, given a new patient record X at the time the patient is discharged from the hospital, can predict the category Y that this patient belongs to. In doing so, they make use of historical data that consists of (i) patient records at the time of discharge from the hospital (many examples of X) and (ii) the categories to which those patients belong (the value of Y for each X). However, a lot of additional data is available about this patient that today's analytics algorithms are unable to leverage. In addition to the data about the patient at discharge time (X), there is also data (call this X') about everything that happened to this patient during the six months after the time she was discharged from the hospital. But the only piece of data that is used after the patient discharge is the six-month outcome: here, this is the category number (one or two).

An obvious question arises: why don't existing analytics algorithms use this additional data? This is because of a fundamental limitation in the way they work. They can only use the data X to learn how to predict Y for a new X; they can't use X', since this data X' is not part of the data that they will have at prediction time.

LUPI, on the other hand, incorporates a breakthrough technique that allows leveraging this additional data X'. We call this data "privileged" information because it is available only during the learning or model formation phase.

The net effect of using this additional data is that LUPI can learn accurately from a much smaller number of patient records than existing analytics algorithms.

More precisely, if existing algorithms need D data records to achieve a certain degree of accuracy, then LUPI only needs  $\sqrt{D}$  data records to achieve the same accuracy. This not only addresses the problem of lack of data for training, but also significantly shortens the required time and resources needed for training. LUPI produces more accurate predictions faster and with lower cost and effort.

## LUPI CAN SOLVE YOUR ANALYTICS PROBLEMS

LUPI can be applied to a variety of analytics tasks where privileged information is available. Some examples include:

- Automated analysis of X-rays to generate cancer diagnosis: Privileged information consists of pathologists' reports, which are available for X-rays that have already been manually analyzed by a pathologist but are not available for new X-rays that must be analyzed automatically to diagnose cancer.
- Anomaly detection: hospital records may need to be analyzed to look for anomalies, such as poor patient outcomes, inefficiencies, etc. Privileged information may be available in the form of expert analysis of past hospital records and annotations of uncovered anomalies.
- Object identification in imagery: low-resolution imagery may be complemented by selectively available high-resolution images for training (privileged information); the latter may not be available at the time when new imagery has to be analyzed.

See [www.peratonlabs.com/machine-learning-data-analytics-research.html](http://www.peratonlabs.com/machine-learning-data-analytics-research.html) or contact us at [info@peratonlabs.com](mailto:info@peratonlabs.com) for more information on LUPI and our capabilities in analytics and machine learning.