



PATTERN COMPUTER®

**THE POWER OF
KNOWING WHY**

Pattern Computer, Inc.
38 Yew Lane
Friday Harbor, WA 98250



Pattern Computer Inc.
Copyright © 2020 Pattern Computer Inc.
All rights reserved.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, mechanical, electronic, photocopying, recording, or otherwise, without prior written permission of Pattern Computer Inc., with the following exceptions: Any person is hereby authorized to store documentation on a single computer or device for personal use only and to print copies of documentation for personal use provided that the documentation contains the Pattern Computer copyright notice.

No licenses, express or implied, are granted with respect to any of the technology described in this document. Pattern Computer Inc. retains all intellectual property rights associated with the technology described in this document. This document is intended to inform about Pattern Computer product offerings and technologies and its implementations.

Pattern Computer Inc.
38 Yew Lane
Friday Harbor, WA 98250

PATTERN COMPUTER MAKES NO WARRANTY OR REPRESENTATION, EITHER EXPRESS OR IMPLIED, WITH RESPECT TO THIS DOCUMENT, ITS QUALITY, ACCURACY, MERCHANTABILITY, OR FITNESS FOR A PARTICULAR PURPOSE. AS A RESULT, THIS DOCUMENT IS PROVIDED "AS IS," AND YOU, THE READER, ARE ASSUMING THE ENTIRE RISK AS TO ITS QUALITY AND ACCURACY.

IN NO EVENT WILL PATTERN COMPUTER BE LIABLE FOR DIRECT, INDIRECT, SPECIAL, INCIDENTAL, OR CONSEQUENTIAL DAMAGES RESULTING FROM ANY DEFECT, ERROR OR INACCURACY IN THIS DOCUMENT, even if advised of the possibility of such damages.

Some jurisdictions do not allow the exclusion of implied warranties or liability, so the above exclusion may not apply to you.

The Power of Knowing Why

Business-Critical Understanding

Gaining insights from machine learning can be an invaluable new tool for a businessperson, technical decision maker, or researcher; it can also present new challenges. While machine-learning algorithms can reasonably predict outcomes when given new observations, as valuable as these predictions are, most algorithms can't tell you the key factors associated with the failure of a manufacturing process, whether your loan approval algorithm is biased, or why the interaction of clinical and genetic factors results in different cancer outcomes.

How do you start to address a problem with hundreds, or even millions, of potential covariates when you don't understand the key factors associated with an outcome? How do you gain insights into higher-order patterns?

We can do it, and we can tell you why.

Machine Learning – Powerful Predictors

Most machine learning today is powered by variations of deep neural networks (DNNs). DNNs can determine whether an image in a crosswalk is that of a stop sign, a yield sign, or a person. In the medical space, a neural network could determine, with a specific accuracy, whether a tumor is benign or malignant, or even predict how long a given patient is likely to survive. In the financial industry, an algorithm can be fed years of loan-approval records to automate the approval process based on the data associated with prior good- and bad-performing loans.

The ability of neural networks to recognize patterns with high accuracy has transformed machine learning and empowered new industries with such capabilities as speech recognition, object identification, autonomous navigation, and reading radiographic images. DNNs can be used as classifiers or as powerful predictors of an outcome.

Neural Networks

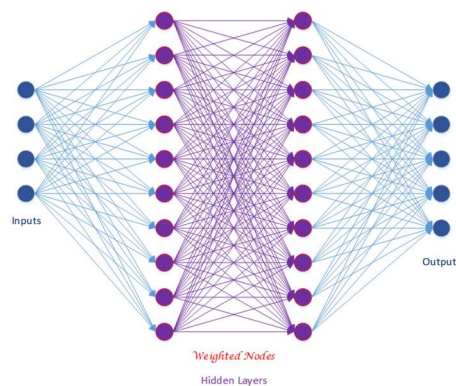


Figure 1: Example Neural Network

A neural network is a computing construct popularized in the 1970s. It comprises a series of layers: an input layer followed – in the case of DNNs – by a number of hidden layers and, finally, an output layer. The input layer consists of nodes based on the number of inputs, while the hidden layers may consist of many nodes based on the design of the neural network. The final layer is defined by the number of different outcomes or classes required for the problem at hand. A user provides a neural network with a very large dataset – in industrial examples, usually tens of thousands to tens of millions of examples. If you are training a neural network to be an automotive classifier, you might need to provide images of cars, trucks, and buses. If you are training a neural network to be a medical binary predictor, you might provide it with genomics datasets of patients with and without a specific type of diagnosed cancer.

Why...?

A neural network tells you the *prediction*. While the result may be accurate, it generally doesn't tell you *why* it reached that result. That is not one of its strengths. The Pattern Discovery Engine™ provides you with a ranked list of the learned features, as well as the correlation of these features to one another. *If you are trying to solve a problem, or want to understand how a complex system works, it is important to understand the factors (working together or independently) that drive an outcome.*

For example, say you are a researcher working to understand the genetic factors associated with disease severity in a specific type of cancer. The Pattern Discovery Engine can provide insights into which genetic factors work together to produce the outcome – not only can it give you the “what,” but also aspects of the “how” and “why” at the same time. With over 24,000 genes of potential interest in the human body, one simply can't scan the hundreds or thousands of patient records necessary to determine the important genes associated with the cancer. It is also becoming clearer to researchers that for a given cancer it is not a case of one gene operating independently, but rather of a combination of genes with different ranges of expressions. The problem is that the brute-force computation of the combinatorial factors for identifying multi-gene relationships (e.g., 2nd, 6th, or nth order) would be very expensive and time-consuming (at best), or simply intractable¹.

Why is “knowing why” important?

The power of the Pattern Discovery Engine is its ability to process large datasets - imagine millions or trillions of columns in a table - and identify the natural patterns within those datasets. In order to understand the power of what the Pattern Discovery Engine can do, we will dive into the realm of medical science, where researchers are working to diagnose breast cancer. The dataset discussed here is based on both a set of measurements taken from tissue samples of breast tumors and whether these tumors were benign (not harmful) or malignant (potentially harmful). In November of 1995, a study was published as the Wisconsin Diagnostic Breast Cancer dataset (WDBC). It is a set of 10 measurements and 20 calculated values based on the measurements taken from the breast tumor tissue sample. (The examples here are very small for ease of explanation.) This is a dataset of 965 breast cancer patients. The technical details of the measurements are here:

[https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+\(Diagnostic\)](https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic))

¹ Consider the factors of $\binom{24000}{6} + \binom{24000}{5} + \binom{24000}{4} + \binom{24000}{3} + \binom{24000}{2}$ and how many years it would take to perform this number of combinatorial computations on a supercomputer.

After running the WDBC dataset through our Pattern Discovery Engine, without any inputs from oncologists or histologists, we discovered that the two most critical features to distinguish between malignant and benign tumors are:

- Worst Area (Column Z) and
- Worst Concave Points (Column AE)²

Indeed, "Worst Area" is a valuable discriminator. Looking at the chart of these data points, as shown in Figure 3, it is much easier to see that the set of values for the malignant tumors is distinctly elevated compared with the set of values for the benign tumors.

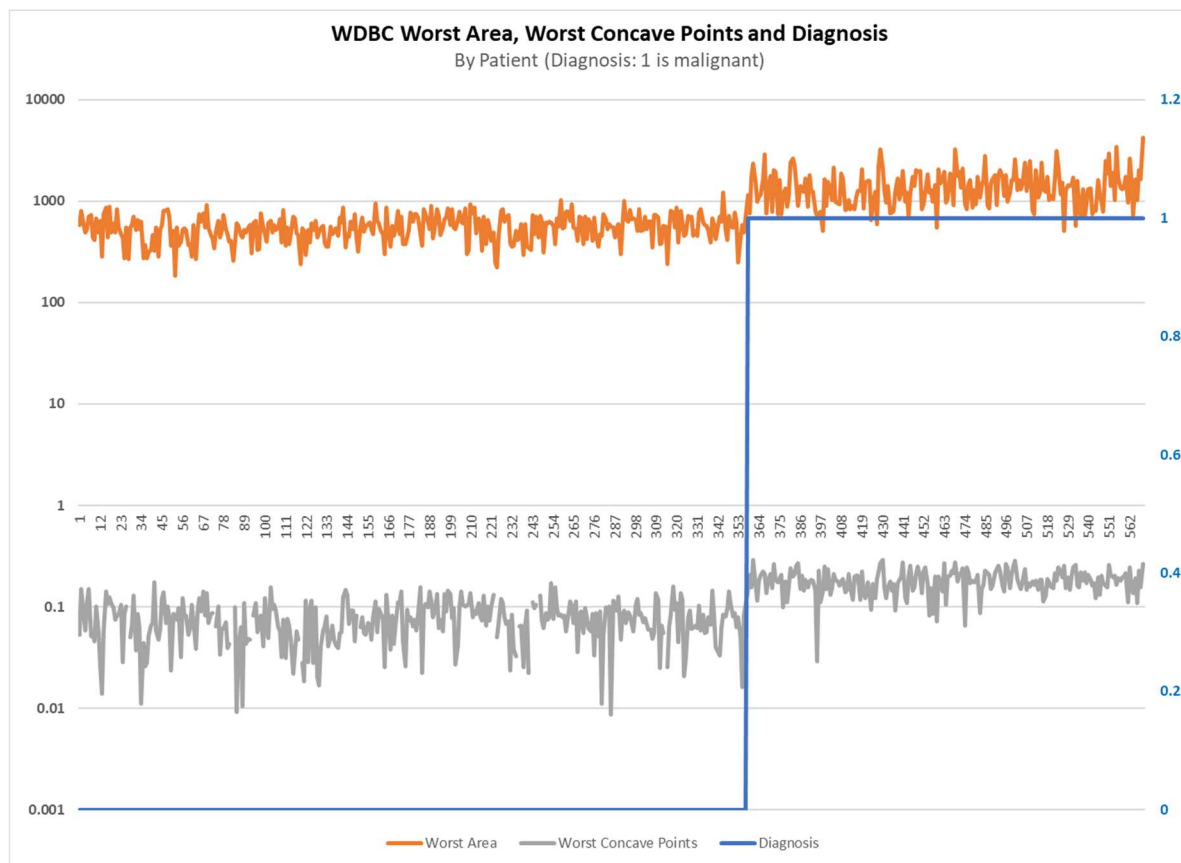


Figure 4: WDBC Worst Area, Worst Concave Points, Diagnosis

² See Appendix A: Sample Data from the Wisconsin Diagnostic Breast Cancer Dataset.

Not only that, but these *two* values can predict with an accuracy of 93.9% – which is easy to visualize and understand:

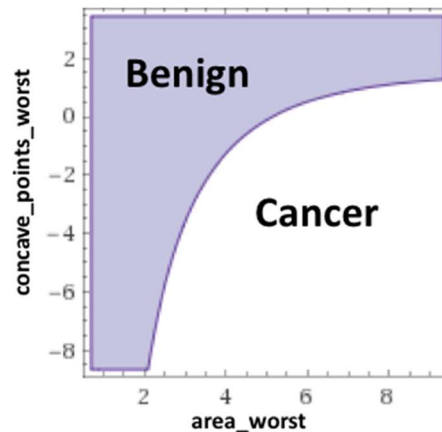


Figure 5: WDBC: Concave Points Worst vs. Area Worst

We know, then, that simply by taking two measurements, we can be over 93% accurate instead of needing the 10 measurements taken as part of the WDBC dataset. If we are able to take only two values, these are the two most important features in the dataset. Moreover, the Pattern Discovery Engine produces the mathematical model which best represents the data with an accuracy of **96.8%**:

```
If(sqrt(texture_worst) * texture_worst * area_worst *  
concave_points_worst) >= 12924 then the tumor is malignant.
```

A Second Example: The Power of Knowing Why

Our second example is based on the heart failure records of 299 patients in Punjab, Pakistan. Collected in 2015, the records contain 13 features³, including whether the patient died during the follow-up period after the heart failure event.⁴ Based on these records, we were able to determine that the serum creatinine and ejection fraction were two of the most critical factors in determining the outcome.

³ Age, anemia, high blood pressure, creatinine phosphokinase (CPK), diabetes, ejection fraction, sex, platelets, serum creatinine, serum sodium, smoking, time, death event.

⁴ Average follow-up period was 130 days.

Using the Pattern Computer Discovery Engine, we quickly identified the *three* critical factors of “**time**,” “**serum creatinine**,” and “**ejection fraction**.” We also created a prediction of **85.3%**⁵ against holdout data with the following model:

```
If ((1.0 / (([time]*[serum_creatinine])) +
([serum_creatinine]/[ejection_fraction]))) >= 0.050998, then it was
85.3% likely that the patient would not live to the end of the
follow-up period.
```

This more accurate view of the survival of heart failure patients could *literally* save lives by identifying patients who are at high risk and need immediate intervention. Instead of having 30 different potential areas to focus on to impact patient survival, you have three specific areas, which can save time, cost, and lives.

Knowing what the key factors are, we can now look at the relationship between these key factors and map out those patients who died and those who survived through the follow-up period, compared with the model, noting the True Positives (TP) (those who the model predicted would die, and indeed died), as well as the True Negatives (TN) (those who the model predicted would survive, and did). The False Positives (FP) and False Negatives (FN) are noted in the 3D model as well.

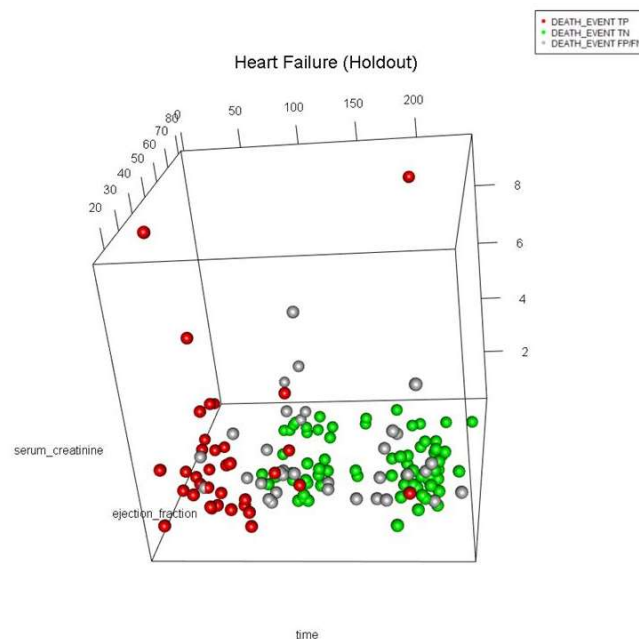


Figure 6: Heart Failure Model (Serum Creatinine vs. Ejection Fraction vs. Time)

⁵ 85.3% accurate true positive results and 85.5% accurate true negative results.

The Importance of Knowing Why

Popular references to “artificial intelligence” are typically used to connote machine learning via the use of neural networks. These algorithms are now being used to reduce the amount of human labor required to make significant decisions for major corporations involved in insurance, finance, manufacturing, rental properties, and the like. The models that they have built are often based on historical data and records of previous good- and bad-performing insurance policies, loan approvals, or rental agreements. Using the historically good- and bad-performing risks in training a neural network may be effective in creating a reliable return for a business, but is that same business owner or decision maker aware of any inherent bias built into these models, based upon the years of previous metrics used to train those models?

With increased scrutiny and public opinion demanding transparency in the decision-making process, knowing what the models are behind your neural-network-based algorithms becomes critically important. What was once a great-performing capability in your business could become an expensive and onerous liability through no explicit intent – but the bias was already in the training data.

Using the Pattern Discovery Engine to analyze the information and results being produced by your neural network, we can discover the key factors being used to make these decisions, as well as the models produced by the neural networks. In this manner, the business owner or decision maker can provide this information to their risk management team to show that there are no inherent biases in the machine-learning-based approach – or if there are, to create new models in which the inherent biases are removed.

It is increasingly important that businesses that currently rely upon machine-learning techniques in the decision making process have a solid understanding of the models upon which these decisions are being based, as they are responsible to their customers and shareholders to comply with ethical and legal requirements of fair business practices. Ignorance is no excuse.

While artificial intelligence may yield above-average performance for your business, the importance of knowing *why* a prediction was made, *how* these machine-learning models work, and the key factors in their decision making is something that any business should understand. Clear understanding of how decisions are made will result in transparency to customers, employees, and managers regarding whether their algorithms are providing unbiased, equal opportunities for employee success and customer satisfaction while keeping true to the company's values and improving performance.

The power of knowing why and understanding the models will be the next trend in these AI-powered industries. As presaged by the GDPR⁶, there are real business risks, with significant penalties, should businesses continue to lack the tools that will enable them to understand,

⁶ General Data Protection Regulation (<https://gdpr-info.eu>)

including being held to account by legal action. The Pattern Computer Discovery Engine is such a tool to assist your business in understanding the patterns and models being used by your AI systems.

In Summary

The ability to understand patterns in datasets allows individuals and businesses to understand the “why” of the revealed pattern: What are the key factors driving the outcomes, and therefore, what action or research should be undertaken to understand these factors that may be leading to undesirable outcomes? As was shown, just being able to predict a specific outcome does not provide information on how to address or prevent an undesired outcome. Knowing what the critical factors are for understanding specific scenarios allows individuals and businesses to focus on gathering just those relevant factors, which can reduce the time, costs, and risks of gathering insignificant information.

A patient dies, a critical business continuity loan is not approved, a company is sued for years of biased decision making: all are potential outcomes of not understanding the key factors informing predictions mechanisms behind AI-supported decisions – the “why.” Acting now to *understand* the decisions being made in your neural networks will be informative as well as save costs and prevent potential damage to your business’s hard-earned reputation. Knowing Why is even more critical to your success than the power of the neural networks themselves.

The power of Knowing Why.

Welcome to Pattern Computer!

