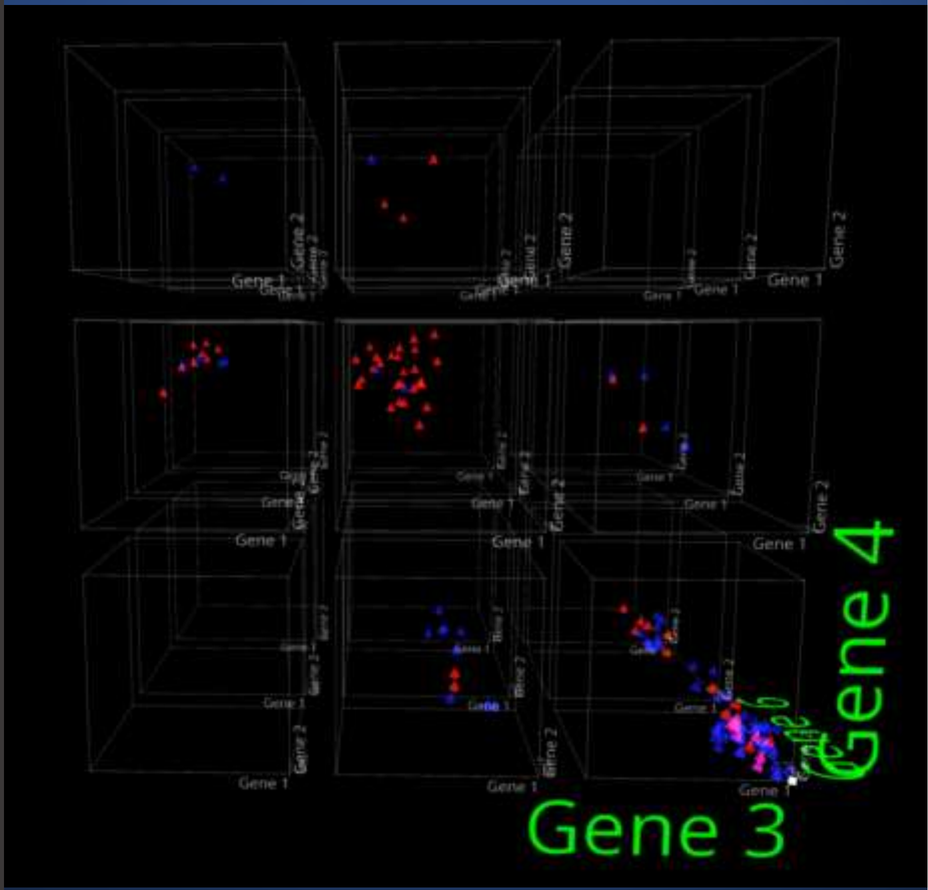
INTRODUCTION TO PATTERN COMPUTER

Pattern Computer Inc.

38 Yew Lane

Friday Harbor, WA 98250



Executive Summary:

Pattern Computer targets and specializes in discovering patterns in complex, high-dimensional datasets. We know that many entities have seemingly endless petabytes of big data in which hidden patterns can be found that capture significant strategic and monetary value for its custodians. But existing machine-learning methods are not particularly useful when applied to discovering interpretable and actionable patterns within datasets. Heretofore, the computational horsepower and/or the need to overcome the mathematical complexity necessary to find patterns within diverse datasets have been insurmountable barriers. Thus, a novel approach must be taken to make previously unseen patterns emerge.

Pattern Computer has solved these problems. Our *supervised* machine-learning approaches allow us to discover and identify the key factors in a dataset most responsible for specific outcomes. We can then use these factors to build mathematical representations of the data to model the relationship. We can also model specific outcomes and make predictions on which factors to change and/or monitor to optimize the outcomes.

Our *unsupervised* machine-learning approaches allow us to identify clusters of similarities within the data. We can identify variables that are redundant, have similar function or representations, and can be investigated for potential substitution – using a less expensive or more readily available component, element, or drug, for example – to accomplish the same goal. We can also identify variables that are complementary or antagonistic, where we know that optimal functioning of a system requires deep coordination between these variables. While our initial focus of development has been on discovering the genomic patterns behind diseases, our tools are dataset-agnostic and have been used to identify issues in energy production emissions, manufacturing defect analysis, and flight departure delays, as well as breast cancer.

While most companies use machine learning (of which the majority are using "black box" neural networks), Pattern Computer has developed its own proprietary algorithms to identify the critical patterns in high-dimensional datasets and is not computationally limited by the $O(n^2)$ scaling issues of neural networks (where n is the number of factors). In addition to pattern discoveries and building accurate mathematical models of the relationships between the factors, Pattern Computer also presents the relationships between these key variables in up to 8 dimensions via our Dimensional Navigator™, which presents the data in virtual reality.

Introduction

Pattern Computer is a five-year-old Seattle-area startup focused on discovering patterns in complex high-dimensional datasets. It comprises a dedicated group of approximately 20 accomplished experts in the areas of advanced mathematics, microbiology, computational algorithms, control systems, applied physics, hardware and software engineering, bioinformatics, data science, computer vision, visualization, and security. Differentiating ourselves from the many machine-learning companies that use neural networks to perform pattern recognition (recognizing patterns that they have been trained to recognize), Pattern Computer uses its Pattern Discovery Engine™ to discover patterns which (of course) have not been seen before. This is compelling because traditional methods that try to recognize patterns take large amounts of compute resources due to the combinatorial mathematics involved in checking all possible permutations and ranking them accordingly.

The Pattern Discovery Engine was designed with advanced, leading-edge mathematics to reduce computational complexity while maintaining the relevant information within the dataset. The mathematics employed do not predicate the dimension or the nature of the results; rather, the strength of the patterns determines the dimension of the results. A 5-way or perhaps 6-way pattern may emerge which allows a subject matter expert (SME) to see the natural dimension of the patterns, versus a researcher who might search with a preconceived outcome and a predetermined number of factors. For example, in cancer datasets, instead of one or two genes, SMEs identified biological pathways associated with poor survivability. In flight operations, they saw flight departure delays due to wildfires on the West Coast of the United States. In energy generation, they saw patterns of temperature, device orientation, airflow, and specific gas concentrations leading to specific types of emissions. Why is Pattern Computer different? Because we discover and rank patterns in the dataset that traditional data analysis tools cannot find.

Once patterns are identified, they are ranked in order of significance. When we used the Pattern Discovery Engine to understand the factors behind flight departure delays, we saw a specific pattern of delays due to the wildfire smoke on the west coast of the United States limiting visibility – particularly ground visibility – during the months of July and August 2018 as having the most significance. The second most significant pattern revealed that icing conditions at the departure airport also has cascading effects throughout the day for the impacted aircraft. Another differentiator for Pattern Computer is that we do not see just one pattern; we see all the patterns in the dataset regarding a specific outcome, in ranked order.

In addition to discovering and ranking patterns, we also build a mathematical model to describe the discovered patterns. In these models, researchers, data scientists, and subject matter experts observe and develop an understanding of how the factors relate to one another to contribute to the specific outcome. How often have you been asked to describe an interaction, or to solve a complex problem, and not had the mathematical relationships to understand which factors may be independent and which interdependent? Neural networks are notorious as “black boxes” whereby the researcher, data scientist, or SME can learn only what resulting prediction is made by the neural network. It is like having an accurate Zoltar machine: the machine may tell you the outcome with some reasonable level of accuracy, but you are unable to understand why. With Pattern Computer, not only are you able to understand the most important factors related to a specific outcome, but you also understand *how* they are related.

Finally, Pattern Computer employs its Dimensional Navigator to be able to visualize up to 8 dimensions using its proprietary virtual-reality environment to understand the nature of the relationships of the different factors with respect to a specific outcome. Just being able to visualize how the key factors are related, and which value ranges are significant (and which are not), creates a new level of understanding: where to focus your limited resources to address a problem.

Working with Pattern Computer

Introductory Discussions

Typical engagements with Pattern Computer start out as a series of discussions in which prospective customers meet with our engagement manager and members of the technical team to discuss the pattern discovery objectives for a potential project. The Pattern Computer team answers questions and may show examples of previous (public) work to demonstrate results from discovery runs, providing examples of the mathematical models and demonstrating how the models compare with held-out data.

The engagement starts with a signed nondisclosure agreement protecting the customer's data and pattern discovery results as well as protecting the work of Pattern Computer. The nature of the specific engagement is then framed out specific to the customer's needs.

We will also discuss the appropriate handling requirements of the dataset. Pattern Computer maintains its data in a secured datacenter with restricted access and multiple layers of physical and electronic data security processes overseen by our Chief Security Officer, Marcus Sachs, who has worked at high levels in multiple US government agencies, including the White House and protecting the United States' national electric grid. In some cases, the custodial requirements of the dataset require that it reside in the customer's own tenancy (in Amazon Web Services, for example) or at the customer's premises (typically in their datacenter). Pattern Computer has anticipated these needs and can make special arrangements for handling such cases.

Data Review

The next step is for the Pattern Computer technical team to review the dataset. This is initially performed by our data scientists, who evaluate the nature of the dataset – whether there are missing values, as well as the nature of the data types (e.g., continuous, ordinal, enumerated, etc.). We will also identify any potential issues that we will need to discuss, such as how to handle missing values in the dataset; and whether we should ignore those observations, average the data, or determine whether there are other specific imputation processes for handling missing values.

Next, our data science team reviews the dataset, the schema, and the joins between the dataset's tables to understand the full scope and interdependencies of the data. Finally, the team reviews the proposed objectives of the pattern discovery process. Some customers will want to run in unsupervised mode, whereby we identify the natural groups or clusters within the dataset(s). Other customers are looking for specific patterns to answer business or research questions and will want to understand patterns behind specific outcomes. As an example, when preparing to run the dataset for US commercial flight operations to understand flight delays, we chose to look at flight departure delays, as they would provide more insight into the operational delays more closely related to the airline and airport operations than would flight arrival delays:

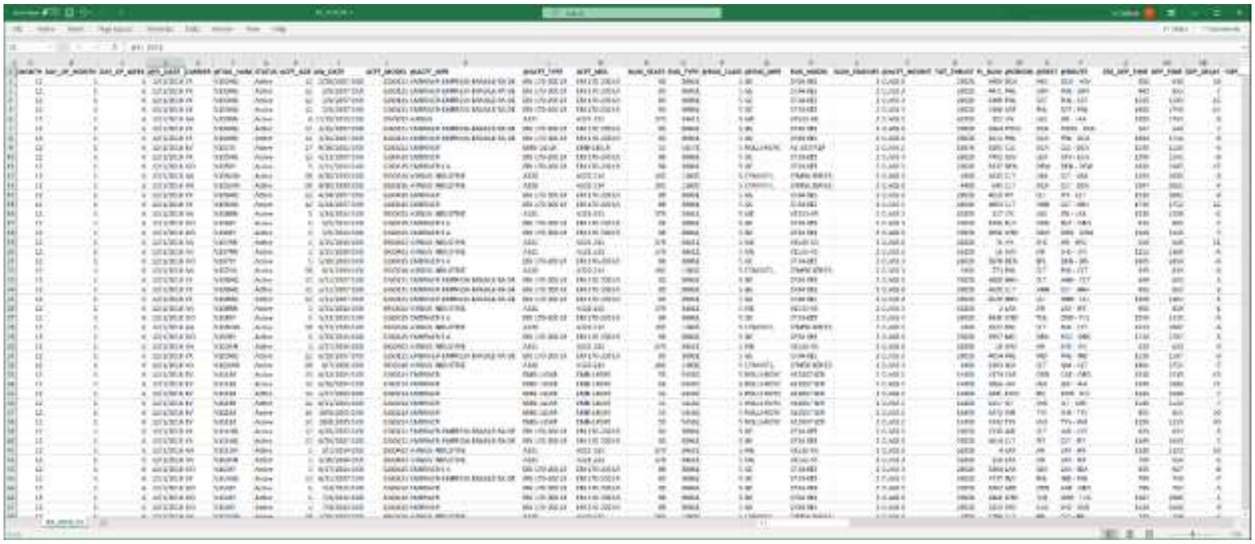


Figure 1: US Flight Operations Dataset - CY2018 (7.2 million flights x 106 factors)

In the case of the breast cancer dataset, we looked at the contrasting data patterns between people who had breast cancer and survived greater than 5 years against those who survived 3 years or fewer.

Working together with the customer's data science team, the customer's subject matter expert, and the Pattern technical team, we will work to understand the customer's goals and propose specific runs to achieve that objective. In general, this determines the target covariate (such as flight departure delay time). For example, the customer might want to know the flight departure delays considering *any* departure delay as an important delay event; by contrast, the FAA only considers a flight delayed if the departure has been delayed 15 minutes or longer. We will want to make sure that we have a good understanding of the specific nature of the customer data as it applies to the outcome(s) we are running the pattern discovery against. Depending on the specific domain of interest, this could be well production, pollution emissions, energy production, salmon stocks, insurance payouts, survival time, or almost anything else. The Pattern Discovery tools are not specific to a given domain or problem. The most important element of getting quality results is having good, clean data to provide as inputs to the discovery engine.

After the Initial Runs

Once a set of initial runs is complete, the Pattern Computer technical team will request a meeting with the customer's data scientists and SMEs to go over the result sets. These initial meetings are fairly revealing to the customer's data scientists; they will likely see patterns that they may already be familiar with, but it is also likely that they will see covariate interactions that were not previously known, or perhaps that the team had suspected but whose role they were not previously able to pin down. It is not uncommon for new patterns to be revealed that were not previously known or understood. Those "Aha!" moments are fun and provide important revelations to work with. Fortunately, the Pattern Discovery Engine tracks the specific observations in the dataset identified by the pattern, thus enabling the customer's data science team and SMEs to dive deeper into that specific subset of data to further understand the driving scenario behind the pattern's existence.

Latent Variables

In some cases, the patterns revealed will indicate one or more missing, or latent, variables. All the input data may be clean, well-structured, and accurate, but the results indicate to the SME that the story being told is perhaps one level abstracted from the real factors related to the specific outcome. For example, when we were running the Flight Operations dataset and looking for the patterns behind flight departure delays, the initial results were given (for the top-ranked pattern) as Flight Date, Tail Number, and Route:

Interaction	Months
Flight_Date, Tail_Num, Route	1-12
Tail_Num, Route, Dep_Time	1, 2, 3, 6, 8, 9, 10, 11, 12
Tail_Num, Dest, Route	2, 3, 4, 5, 10, 11
Tail_Num, Dest, Wx_Arr_CldCvr	7, 8

Figure 2: Flight Operations: Initial Results - Flight departure delays

Stated plainly, the pattern was saying that the most significant reason for flight departure delays was related to a specific set of aircraft (as identified by their tail numbers) that were significantly delayed on specific dates and on specific routes. But what the pattern was not saying was also very important. The dataset included the weather at the departure airports at the scheduled departure hour, as well as at the arrival airport at the scheduled time of arrival. None of those factors were identified in the pattern. The most significant factor behind flight departure delays was not about local weather events at the arrival or departure airports.

The SME loaded the dataset and observed the planes most impacted by delays, then looked at the dates involved and the routes those aircraft were taking. Accustomed to looking at aircraft and airport three-letter codes, the SME quickly observed that these impacted flights were transiting into the New England area on a set of dates during which the longest delays occurred. The planes most impacted were on shorter routes that transited the area west of the New England area, and they were doing a series of out-and-back routes, or circular routes, in that region. The specific airports did not report severe weather situations, but it was clear that something was seriously impacting the flight departures.

There is another area of the FAA that plays an important role in the flow of aircraft: the FAA Operational Network (OPSNET), which manages the flow of aircraft across the country using the available jet routes. Using Traffic Management Initiatives, they can delay flight departures by not granting clearance for takeoff at the departure airport if there are no available jet routes to handle that aircraft due to air traffic volume, en route weather, or other factors. The SME looked up the weather west of New England on those dates and saw a line of massive storms all the way from the Ohio Valley south to Texas. The missing dataset was the TMI information from the FAA indicating flight delays due to weather and volume:

Interaction	Months
Tower_To, Sched_Dep_Time, Leg, Tower_From	1-12

Figure 3: Flight Operations - With the FAA Operational Network TMIs added (due to weather)

Once that additional dataset was overlaid on the Flight Operations dataset, we ran the Pattern Discovery Engine and found that the flight departure delays were primarily due to FAA OPSNET delays attributed to "Traffic Management Initiatives due to Weather."[1] Seventy-five percent of these delays were initiated by the destination airport tower not giving clearance for the aircraft to depart. That dataset included the missing covariate information.

Mathematical Models

Along with identifying the most important factors in a dataset, the Pattern Discovery Engine also produces an accurate mathematical predictive model of the given data to classify against the specific outcome. This information is very useful in understanding which factors are showing up independently in the models, which are correlating toward a given outcome, and their level of contribution toward that outcome. Additionally, close inspection may reveal specific clusters of factors responsible for a given outcome only under certain conditions described by other factors. The model provides powerful insight into the nature of the datasets and interactions or correlations among the factors.

In some cases, the mathematical model can exhibit useful parsimony. For example, the Wisconsin Breast Cancer Dataset[2] includes 30 factors based upon 10 patient measurements related to breast cancer. The mathematical model (in an Excel formula), which demonstrated 99.1% accuracy on held-out data, was simply:

```
= IF(smoothness_worst * area_worst >= 104.44413, 1, 0)
```

This very simple equation, which notes that only two of the 30 factors are needed to make an accurate prediction regarding the presence of breast cancer, requires that only two patient measurements be taken, thus reducing time and cost associated with gathering the data necessary for an accurate diagnosis based on data. Notably, this model was more accurate than that in the published paper associated with the dataset.

Dimensional Navigator™

It has been said that seeing is believing, but in the data science arena, seeing relationships enables understanding. Pattern Computer has developed the Dimensional Navigator, which allows customers to visualize the relationships between the critical factors and visually map their relationships in up to an 8-dimensional space in virtual reality (VR). In addition to navigating the relationships between the factors, the Dimensional Navigator allows the user to query the data values of the individual objects in real time, allowing them to step through the data and the related values one at a time in the VR context.

[1] "Tower_To" and "Tower_From" are the FAA Operational Terms for TMIs initiated by the destination and departure tower controllers, respectively. These TMIs were specifically caused by weather events.
[2] https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

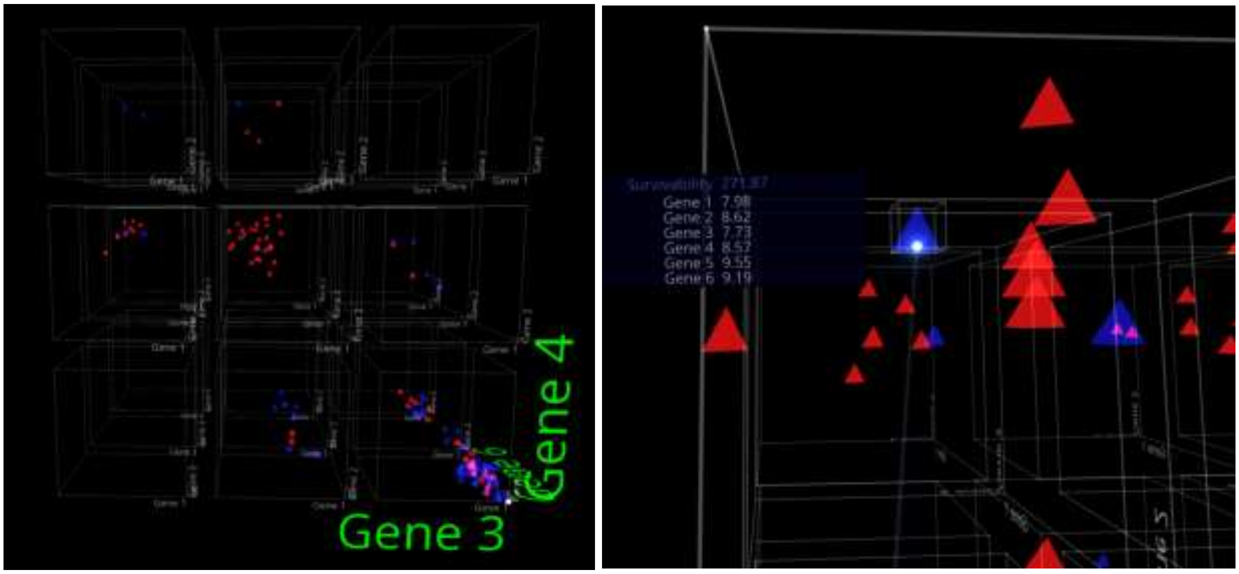


Figure 4: Views of the Dimensional Navigator hypercube representation (L). Querying values of the data object (R).

Pattern Computer continues to develop additional capabilities, driven by both customer requests and the company's strategic plan to dive deeper in the areas of causal relationships and more advanced mathematical capabilities. In addition, we are expanding our breadth in providing customizations for specific domains, developing domain-specific enhancements based on the nature of the pattern discoveries identified.

Summary

Pattern Computer is a Seattle-based startup focused on discovering patterns in complex, high-dimensional datasets. It is different because it:

- Discovers patterns versus just recognizing previously known patterns (the dimensions of the patterns are not predetermined)
- Recognizes and ranks multiple patterns in the datasets
- Supports very high-dimension datasets (at least tens of millions of factors)
- Builds accurate models representing the patterns (as measured against held-out data)
- Creates up to 8-dimensional virtual-reality models of the relationships in a pattern
- Partners with your data science team and researchers to discover the patterns and insights

Pattern discovery is the way to distill the important information and capture the value from large, complex datasets in cases where brute-force combinatorial analysis is simply not cost-effective or computationally tractable. We have already experienced breakthrough-class discoveries using these powerful tools. The toolset is not specific to any given domain, having been applied in fields including genomics, biology, aerospace, energy, and aviation, to name a few. Pattern Computer can run the dataset in its secure datacenter but can also make special arrangements to run in the customer's cloud environment or remote datacenter. Future work will expose the Pattern Discovery Engine as a web service.

Frequently Asked Questions

Why is Pattern Computer the most advanced, globally?

Pattern Computer can discover patterns in high-dimensional data and identify the features most responsible for specific outcomes. This ability to understand and explain the model in high-dimensional datasets is truly the definition of Explainable Artificial Intelligence, or XAI. There are a few other companies now arriving on the scene starting to claim that they have similar capabilities, but it is important to note that they are limited in handling datasets with more than 3 dimensions, as their techniques are based on more general topological approaches to understanding data. As noted above, the computational complexity issues and advanced mathematics required to make this jump require both mathematical and computational agility, designing computational systems matched to both the hardware capabilities and the mathematical algorithms to produce results in short timeframes (days), as opposed to weeks or months.

Other companies in our space are not producing accurate mathematical models to represent the patterns that thus enable researchers, data scientists, and SMEs to understand the nature of the correlations. The insights gained from these models create the opportunities for new or expanded understandings of the nature of the system.

Our algorithms also can identify missing or latent variables in the dataset – which are then surfaced to the subject matter expert with the related characteristics to assist in identifying the missing data to add to the computational pattern-discovery runs. Finally, other companies are not taking the findings of the algorithms to see in an integrated space, visualizing the relationships within the dataset displaying the data up to 8 dimensions presented simultaneously.

Why would organizations want to be part of Pattern Computer?

Machine learning has shown that it has the ability to perform pattern recognition and to carry that performance to new datasets once it has been trained with thousands of training examples. Pattern Computer builds on top of aspects of machine learning and applies its own mathematical techniques to implement pattern discovery algorithms and extract actionable mathematical relationships. Through this work, Pattern Computer has been employed to discover patterns behind cancers and other diseases – to gain new insights, for example, for disrupting cancer pathways and revealing new understandings of the interactions and associations to cancer researchers and bioinformaticians. Beyond the obvious benefits to the human condition through understanding diseases, the Pattern Discovery Engine has been used to identify methods to minimize toxic emissions from power-plant operations, allowing power-plant operators to respond more quickly to rapidly changing supply-and-demand curves to interoperate in a world of renewable energy supplies, creating more reliable, more efficient, and less polluting power management.

We live in a world of big data. Corporations have been collecting petabytes of information over decades of operations, but few have been able to turn their big data into tangible business and strategic assets. One of the greatest challenges is moving beyond the limited statistical models to gain a true understanding of the data and of the key factors associated with specific positive or negative outcomes (for example, profits and losses). These big datasets have become unwieldy by traditional computational methods, requiring supercomputer-class computers to discover patterns using brute-force methods that are not cost-effective for the companies. New techniques must be used to capture the strategic insights and monetary opportunities locked away in these datasets.

Pattern Computer creates the tools and opportunities to provide social good through new insights into the microbiology of diseases and identification of the key patterns behind them, and to find patterns in toxic emissions and environmental risks. By realizing the patterns in a corporation's big data, Pattern Computer can monetize the value of the big data sleeping on hard drives locked away in datacenters around the world. This may be in the form of new insights in understanding customers' purchasing motivations, discovering patterns in manufacturing defects based on manufacturing test/sensor data which maps directly to parts that will fail prematurely when in service, or discovering patterns behind fraudulent insurance or warranty claims.

How is your capability unique and potentially disruptive?

Our capability is unique and potentially disruptive in that the Pattern Discovery Engine:

- Is the only known set of developed and deployed algorithms to discover patterns in large, high-dimensional datasets. Using the Pattern Discovery Engine, we specifically identify and rank the essential covariates in the datasets that are directly related to the specific outcome, such as whether: a patient has cancer or not, or a part will or will not fail before its warranty period expires, or a loan application is or is not a good risk for the lender.
- Uses the essential factors and builds an accurate mathematical model of the relationship of these factors to the specific outcome.
- Identifies the nature of the relationships within the factors themselves – linear and nonlinear relationships, inverse and direct relationships, etc. The nature of these relationships informs the SME under what conditions these patterns are created, thus telling a story to the SME, who knows how to meaningfully interpret the combined information.
- Identifies the samples (rows) within the dataset that contribute and align to the model. We let the data tell the story.
- Reveals the sub-patterns and the cluster(s) within the dataset that align to the sub-patterns.
- When in unsupervised mode, identifies clusters within the dataset.
- Identifies the factors in the dataset that are most similar to one another. This is particularly useful in the biology domain, where we can identify similar covariates for pathway identification and reduction.
- Can identify the decisions made within a neural network, thus opening the heretofore "black box" of a neural network.

The collection of the Pattern Discovery Engine's capabilities is uniquely disruptive in the space of biology – particularly regarding gene expression. This combination may be disruptive in other domains as well.

Who are you comparing against?

Who do you consider to be your peers?

- Companies such as Palantir, IBM Watson, and the machine-learning cloud services offered by Google and Microsoft.

Are you positioning to lead in modernized domains, or just to help modernize non-modernized domains?

- We are positioned to lead in both. (Perhaps the one modernized domain that is oversaturated in this space is high-speed transactional trading and arbitrage, so it is unlikely that we will spend time in that space.)

How does your capability compare with that of said peers?

- We are not aware of any company currently making claims of discovering patterns in large, high-dimensional datasets and producing explainable mathematical models of same.
- IBM Watson makes claims in the machine-learning space related to pattern recognition, and many companies make claims regarding pattern recognition and imaging capabilities that are based on pattern recognition, including natural language processing (such as Alexa or Siri) or automobile navigation (such as Tesla or Mobileye). While Pattern Computer also has pattern and imaging capabilities, we are not making unique claims in that space. As we mature our imaging capabilities, we will be able to combine our pattern discovery capabilities with our imaging capabilities to advance and broaden the scope of our work, which we expect will also be industry-leading and unique as a combined solution.

What are your most impressive claims?

- Pattern Computer first started in the biology domain. One of our first projects was related to identifying a breast cancer gene therapy based on gene expression data. We used the METABRIC dataset of gene expression data to identify the prominent gene expressions related to poor survivability in triple-negative breast cancer (TNBC). Most studies related to cancers focus on a single gene's relationship to a cancer – at most, two genes. Instead of predetermining how many genes may or may not be involved in poor outcomes, we let the data inform us regarding the dimension of the gene expression anomalies and potential interactions regarding poor survivability. The results were identifiable as biological pathways by subject matter experts. As a result, we were able to create gene therapy "cocktails" based on known, approved drugs. Two of these cocktails have passed beyond *in vitro* trials and are currently in mouse trials, with the goal of stopping the growth of cancer in laboratory mice with TNBC. METABRIC data has been available since 2012, but as yet no one has identified a successful gene therapy for TNBC. Currently, Pattern Computer has successfully completed four series of laboratory tests on TNBC.
- In response to customer questions regarding how our tool compares with published, peer-reviewed academic papers, we have created a short whitepaper, "**Pattern Discovery Beyond Published Results: Three Papers**," which provides specific responses to this question.

Do you have open-source repos, with documented results?

- We do not plan to document a repo for others to be able to reproduce our results. The mathematical models, when applied to the dataset, will produce the documented results. Our models speak for themselves.
- As a commercial startup, we do not make open-source repos available. We are a for-profit company and have stockholder expectations to meet. We have a team of experienced PhDs in various complementary disciplines working to advance the state-of-the-art and excel beyond any published works. The models can be run against the datasets to produce our results, reflecting our uniqueness in this space.

Do you have competition results?

- To date, we have not participated in Kaggle competitions, particularly as some require that we include the code necessary to run the algorithm that we used to produce the results – even though we can produce a model that at least matches, if not exceeds, competition results.
- As a startup, we must be cautious about when to reveal to a broad audience that we take a different approach to machine learning. There will be a time when potential customers will be able to run limited datasets through the Pattern Discovery Engine to build a model and demonstrate our capability. We expect that when Kaggle competitors produce top results using our Pattern Discovery Engine it will be noted by the machine-learning community.