A Practical Perspective of Applied Artificial Intelligence for Precision Based Medicine

By Mike Baciewicz

Progress in the areas of genomic mapping had generated considerable excitement of finally being able to use the genetics as an analytical predictive platform to anticipate and address diseases of the future. The prospect of diagnostic specification for epidemiology has result in significant investments of funds and resources that unfortunately have not translated into visible progress towards an efficacious system for healthcare improvement or disease avoidance. This paper is intended to address the baffling lack of headway, challenge of constructs(s) identification, disconnected overlays of disparate databases, and possible roadmap to translate genomic innovations into a roadmap of future disease predictability and prevention with integrated, Artificial Intelligence staging.

Upon meeting with and collaborating with various businesses dealing with cancer research and treatment approaches it became clear that the specific industry of Applied Precision Medicine is nascent and searching for better tools to expose and correlate the vast, unexplored universe of available data. The industry of Applied Precision Medicine does not contest this and to the contrary publishes this openly and willingly. The simple fact of the matter is the ability to handle data, and handle that data at high volumes and high speeds is here. However, through all this time of tech-advancements in so many areas the root need of better problem-solving has been left in its archaic form with its base approaches formed in the 15th and 16th centuries. The current outcome of that combination yields our world fantastic volumes of data thrown against the wall of hope, with hundreds of bilions spent on it, and yet coming up with the same lethargic outcomes; albeit at a much faster pace.

In reviewing what appears to be the generally accepted norms for data sciences as applied to epidemiology, it appears we are creating vast piles of disappointing results. However, these disappointing results continue to generate enough hope that millions and millions of research dollars get approved to continue to find even marginal advances since the long-term survivability of humans is at stake.

Taking a different perspective and approach, we believe that the current mechanisms in place might have their place but in a far more limited capacity than currently assigned and revered. With regard to approach, we are of the position that the areas under treatment (epidemiology) and the application of currently available technology (ours) can be applied with significantly greater correlated outcomes of parameterized and non-parameterized data.

We believe that a greater depth of applied advanced analytics, with a far larger umbrella of networked data can be mined and correlated with materially greater volumes of useable relationships exposed at an exponential rate.

First, we must address some of the basic approaches taken to address the properties of high-dimensional data spaces as it relates to implications for exploring genomic and protein expression data, and where our opinions might differ from prevailing thought. We will use existing methods and

structures and attempt to stay within the vernacular of the trade as we attempt to discern where today's methods could be improved upon using techniques more common to process engineering and control.

We recognize and respect the inverted relationships of data sets being enormous and subjects being limited, yielding many challenges as it pertains to correlating multimodal, high dimension data with a paltry but growing collection of diverse, disconnected data stores of DNA, blood work and fecal data, to name a few. Further, in the United States, data hoarding is still far more prevalent than data sharing as privacy and financial goals still reign over ethical goals. So the systems we design and build must recognize and have faith that over time the data hoarding will ease, and as the slope of that line decreases we will be met with the great fortune of dramatic increases in the slope of the line for epidemiological success stories. Once that crucial relationship is characterized and shouted across the planet, there will be much greater energy behind responsibly managed and totally shared medical data. We will leave out the levels of de-identification that will likely be instituted, and instead be mindful in our designs that if our investors and donors think first order, single pass methods should be solving or mitigating the human death cycle we should remind them that the current rate of success is the current outcome, which is severely lacking in magnitude of performance. However, much of the infrastructure is in place, it's just not connected correctly and in desperate need of significantly more similar, distributed correlating engines on top of, inside of and around the existing infrastructures of medical data, raw and reduced.

To that end, we can treat a number of the current analytic approaches to gene and protein expression data as tools with varying degrees of efficacy, with that efficacy still largely undetermined due to the current rate and availability of data. This is true across many cancer-research fronts, with challenges such as limited access to DNA so great that trying to convince those with the power that we really need series of DNA tests for each person to really triangulate on the multi-order challenges of "layered, multimode, high dimension data". We are trying to watch a movie and predict who the villain(s) are, but we currently get to look at only 1 frame. We must convince those with power to find large groups of people possessing DNA samples of any kind and institute additional high-quality, data-rich DNA tests at intervals in their life so that we can begin to see a primitive "flicker-version" of the movie we so desperately need to watch.

While those efforts need to be underway and getting funded, we can turn our attention back to reality. Keep in mind our objective is to build a system today that has the ability to ingest any known high-to-medium value data that has some characteristics that can have epidemiological value. We must then go through that data, collecting and storing any outcomes that can show relationships or patterns. And we must do this sequencing changes, sequencing magnitudes of values on currently known relationships, and processing the known relationships for even greater efficacy on additional, heretofore unknown spaces in the environments under review. And we must be patient. Too often it appears we have declared the total outcome value in a space of time rather than a space of potential value. We must think of the process as a lock with an unknown number of ganged and networked tumblers. And those ganged and networked tumblers are also ganged and networked, with multimodal influences with varying magnitudes, based on circumstances that change but we have only a snapshot of.

If one can embrace this perspective, one can better make the jump to comprehending that we are building, ever so slowly, one of the largest disparate relational databases ever known: how to keep a

human alive, optimally, and predictively. To that end the current approaches have myopic designs that are so use-case specific and API-centric that each micro-project's beginnings likely do more to lock out incorporating newly available data sets, rather than design in capacity for adding new repositories of data with value.

Our base software model (The Knowledge Molecule®) happens to be molecular in its base data structures, which allows for greater conceptual assimilation of new/future data sets into designs that can be built now in this space. Additionally, by "stepping back/panning out" from the current holistic view of how this research is done with software, we believe strongly that instead of forcing currently used concepts onto the available disconnected data sets and "seeing what happens", our analytics have a designed-in logic using a Doppler Effect for auto-searching for relationship-associations in the available noisy, disparate data environment.

In conjunction with our patent pending engines, we believe the current methods employed for analytics fall short of the true potential to find correlated relationships in gene/protein expressions. To wit, we disagree that the currently held thoughts on the limitations of the following methods can be relaxed if we accept the "tumbler method" for building the test case of a super-molecule, super-correlated data lake of disparate data.

With current notions of "Data clustering," for example, we must not force a hard or soft clustering structure assignment approach as this is too restrictive, as that approach is human-simplified, where the realm of all possible data associations deserve the possible best-outcome opportunities coalesce naturally (perhaps combinatory transpositional integrated overlays, for example). We must let the logic determine what the relationships are, and the magnitudes of influence in these relationships. And any modality influences will reveal themselves over time, as data-set volume, richness and noisiness grow with new data-set additions. Additionally, the logic and associated analytics must have the capacity to recognize via Artificial Intelligence (AI) when a newly attached database has use-case elements for in situ databases with in situ applied analytics. In this manner, the human element does not have to be cognizant of the potential use/value-case of a newly acquired database. The application will find newly attached or newly realized data-relationships and suggest associations either A. For review and acceptance into the processing cycles or B. Automatically process any found legal datasets for new high-value correlations (which can be circuitous, many "legged" and well beyond current human-construct-recognition). We must accept that these can and will be 1st, 2nd and 3rd order deep dives into the data likely well-beyond our human capacity to see the multivariate, multimode associations, so we must allow them to be generated and auto-vetted systemically until the efficacy tables reveal an outcome that is approaching acceptable levels of consideration. We can do this with our solutions so that the efficacy tables can be watched as the enormous tables of potential associations self-build, ever approaching the threshold of user-selectable value predictions.

This is where we significantly diverge from known analytic approaches. We believe the patience coefficients and data-association engine's complexities are both far below the needs to allow for inspiring confidence in the long-term buildup of data structures that have immediate and long-term value inherent in the system build. And due to that shortcoming, much of todays available data can be coming back as over-fit, not characterizable, and passed over as valueless when in reality it was our collective shortsightedness that had us see that one frame of film, not see anything of interest, discard or ignore the momentary data hence moving forward without keeping snapshots of that frame so that

we can be prepared for when our AI engines begin to recognize that another frame of the same movie just got created 3 databases away.

With this notion, imagine how many times we have likely parsed available data and tossed it on the floor well prior to having advanced AI engines hold the information while waiting to see what dendritic connections, under what circumstances, with what influences and their magnitudes were extracted that have build-up potential for forming the critical mass of a growing-efficacy value-outcome molecule for epidemiology. Predicting diseases and life limiting maladies should not be science fiction.