Human-Centered AI for BI Industry

LF AI & DATA

BI & AI

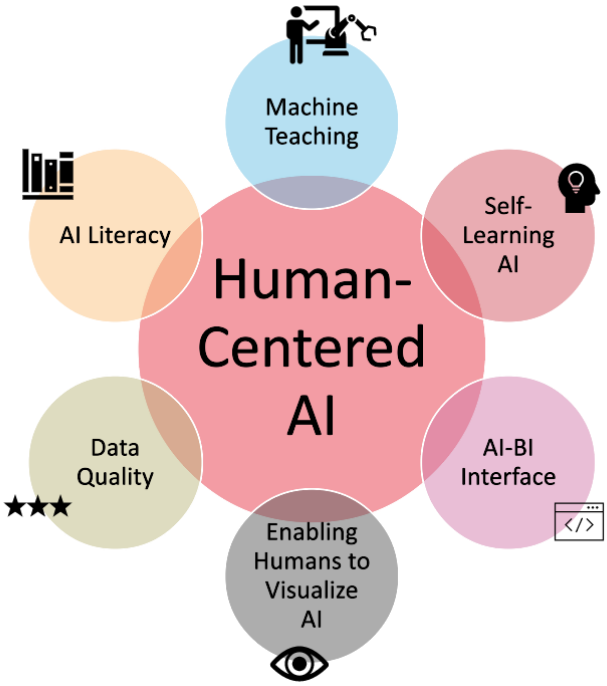
December 2020

Introduction by Cupid Chan, BI & AI Committee Chair

When we want to "talk" to a database, the language we need to know called Structured Query Language (SQL). It is, as its name suggested, very structured and from a functional perspective, the language fulfills its purpose pretty well as we can tell the machine to execute our instructions. However, when you look at this from a user-friendly perspective, even professionals may sometimes have trouble to understand the intricacy of the language because it's not designed with user experience in mind when it was first launched in 1970s. That's why Business Intelligence (BI) tools was born and became the translator by introducing an interface on top of SQL to allow better human interaction. BI converts high level request to a low-level SQL language for a database to execute. This shifts the paradigm from computer-centric to a more human-centered approach for the end users.

History repeats itself. Today, we have another set of "languages" because of the new wave of Artificial Intelligence (AI). Even though a lot more people are diving in to learn these languages and the popularity grows at a much faster rate than when SQL was introduced, the very same gap of computer-centric VS human-centered is still here. Just like any other technology, and AI is no exception: Unless we understand we, human, are the ultimate beneficiary, technology is just a toy for a geek. Fortunately, it doesn't take long for the AI community to realize this gap. For example, Stanford University started a Human-Centered Artificial Intelligence institute led by Fei-Fei Li and John Etchemendy last year. Human-Centered AI will for sure be a trend for the AI community. This should also set the direction to let AI be truly beneficial to our society.

In 2019, our group explored how BI is being impacted by and should respond to the AI phenomenon. This year, BI & AI Committee takes a step further to investigate this influential topic of Human-Centered AI. A group of BI and Analytics leaders dissect this subject into six different areas to see how BI industry should adopt to this important theme.



Machine Teaching by Cupid Chan

Even though the term Artificial Intelligence (AI) was coined back in 1956 in The Dartmouth Conference, the true golden era of AI started in 2012 when Jeff Dean and Andrew Ng published their paper Building High-Level Features Using Large Scale Unsupervised Learning leveraging multilayer neural nets known as deep neural networks, a subset of Machine Learning (ML). Since then, a lot researches were done on how ML can handle different challenges by various algorithms. Moreover, most of the result were shared in open-sourced libraries and frameworks such as TensorFlow and Scikit-learn. We can then bring our own data and take advantages of the optimization already done on the algorithm by other data scientists. Furthermore, with transfer learning, we can even take the result of pre-trained models and reuse that to the new data set as a starting point to cut down a lot of training time. What a leap in ML in just past 8 years.

In my YouTube video "A.Iron Chef" (https://www.linkedin.com/pulse/airon-chef-cook-up-ai-your-data-kitchen-cupid-chan/, https://www.youtube.com/watch?v=xifaCq4F7bg), I summarize 7 steps in machines learning process to show how is the AI/ML relevant to the cooking. As you can see from the diagram below, even though data scientists drive the sequential ML process, algorithm optimalization and efficient model is the focal point with machine learning in the center. Other human interactions are just sparing on different steps as separated activities to aid the "learning" process. For example, domain experts use brute force to annotate data as the input of the learning. Once this is done, data scientists pick it up and continue the ML process. Until the model is finalized and deployed to Production, the end user will then get the result decided by the machine

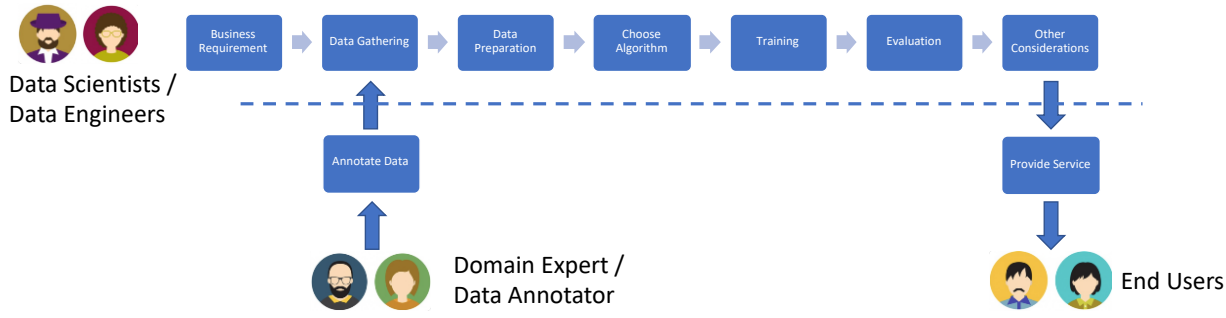


Figure 1: AI Recipe for Data Kitchen

Just like a successful education system, learning is only part of the contributing factor. What about the other side of the equation – Machine Teaching? AI/ML, without a busines context to anchor on with the purpose of serving human, is just a toy for a geek. Yes, it may be fun to play with but there will be a disconnect to the beneficiary – we human. In order to turn this around, we need to inject human into the overall picture and put human in the center of this picture.

Input

Data is the ingredient of all ML Supervised Learning recipes. In order to get it right, human uses brute force and follows a lengthy, error-prone, and tedious process to annotate data traditionally. In order to properly inject human into a teaching process, we let the machine to read the partially annotated input data by itself. At the same time, we allow the machine to be uncertain when it encounters something that cannot be deduced based on the previous annotated data set. What is "uncertain" you may ask. This is the threshold we can set depending on the criticality of having a correct label. If a machine comes up with a confidence score below the threshold, it will ask the human for clarification. Human will then respond, not only with the label, but potentially with the reason why it should be labelled that way. That means the reasoning will be taught, instead of just letting the algorithm to learn by itself.

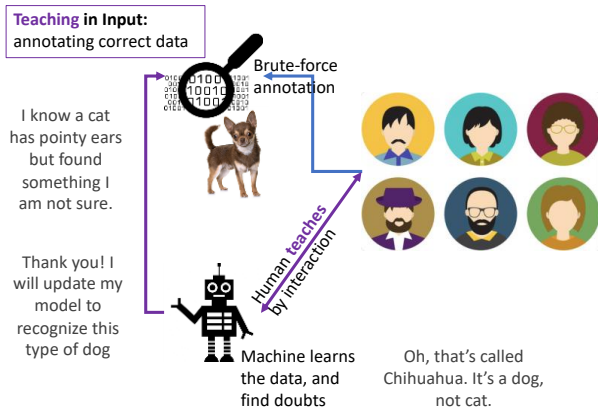


Figure 2: Machine Teaching in Data Input Process

In BI, vendors can incorporate this approach to replace the traditional human data wrangling feature. The tool can explore the data provided by the users and learn the preliminary patterns demonstrated by the data set in an unsupervised manner. Only when the machine encounters something that does not make sense, it will use natural language to query the user. Even though this approach may miss a few data points because human is not exhaustively walk through each and every observation, this on-demand interaction between human and machine allows a much better scalability of annotation. The reliability will also be increased as human engages more in the teaching with a reason, yet not using brute force to walk through millions of data points.

Output

On the other hand, when a model is built and deployed to a production system, we should not just accept that as the final decision to provide the service, which is static and not reflecting relevant considerations. Similar to what we do for the input, we allow the machine to return a "state of uncertainty" when it falls below certain threshold defined by the users. And when this happens, human step in and teach the machine why certain decision should be made. At the same time, we can also perform audit to the result to "accept" or "reject" the prediction generated by the trained model. This human feedback will

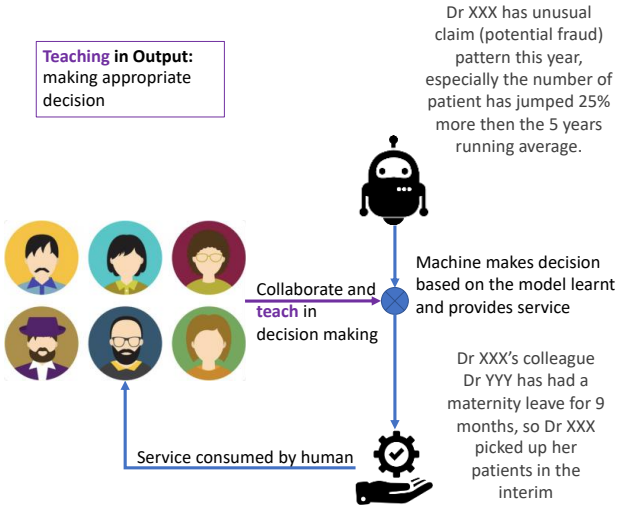


Figure 3: Machine Teaching in Model Decision Process

then be recorded and set back to the system as the input for another round of training. As we continue this iterative process, we teach the machine by providing dynamic feedback. More importantly, we convert the one-way machine learning, to bi-directional machine learning and teaching process.

Most modern BI tools provide ways for users to insert comments for the sake of communicating and collaborating among users. We can evolve this feature to another level by making a dashboard showing not only the predicted results of a model, but also feedback fields allowing users to enter confirmation (accept/reject) and comments of the predicted result. The comments provided by the users can teach the machine by feeding that into a natural language processing (NLP) engine. Hence the original data will not be the only input of an evolving model. By combing the original data set with the users' feedback will make the cycle completed by machine teaching.

From AI to IA

The following diagram summarizes the overall Machine Teaching in a Human-Centered AI. We are all be fascinated by cool technology. That's why Artificial Intelligence have grown and

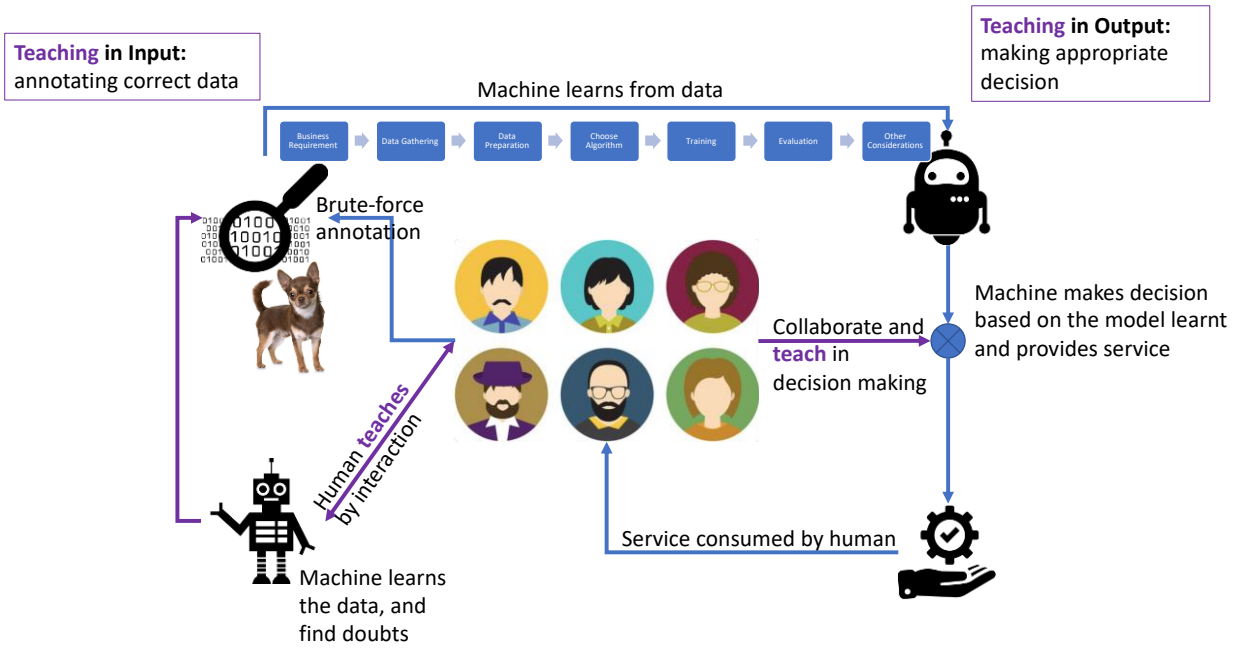


Figure 4: Putting Together of overall Machine Teaching in Human-Centered AI

surged in the past few years. However, putting technology in the center exposes a risk of spending resources without actual benefits experienced by the end users. Therefore, let's do a mindset change and flip AI to IA – Intelligence Augmentation. Putting human at the center, AI, yet another technology, is just peripheral to augment the intelligence of us - human.

The Self-learning AI Engine by Xiangxiang Meng

A human-centered AI requires the underlying AI engine to be as smart as possible. This means we want a human-centered AI that minimize human input or invention in the process of preparing the data, training the models, and select the champion models to generate insights for the BI frontend. Self-learning is the path to continuously evolve the AI engine to get rid of most of routine process from the end user. This includes the capabilities to self-learn to be adaptive to the BI frontend, the platform that hosts the AI engine, and the input problem from the end user.

Rule #1 Frontend Agnostic

The first level of the self-learning perspective of the AI engine is frontend-agnostic. In a human-centered design of the AI engine, users are looking for an autonomous AI layer that can automatically identify and adapt to the BI frontend.

For example, the AI engine should be able to automatically identify which type of BI front-end is sending the request and identify the current version of the BI frontend. Such information can be used to trigger an automatic data wrangling process. In such process, the input data is converted on a column-by-column based (if necessary) by matching the data types used by the BI frontend and the data types used by the AI backend.

Also, the AI engine should learn to automate the data cleanup process. Without extra information from the BI backend, the AI engine should identify data quality checking and data cleanup policy for each column of the input data set. Examples include removing duplicate columns, identify and prevent using ID variables or high-cardinality variables in the models, discarding columns with too many missing values and imputing other columns with missing values.

Rule #2 Platform Agnostic

A self-learned AI engine is platform agnostic. With the recent advances in machine learning, deep learning and reinforcement learning, an AI engine is often an ecosystem consisted of dozens or even hundreds of machine learning, deep learning and other analytics packages and algorithms. In order to achieve the goal of human-centered AI, the engine should self-learn to organize all these packages and be platform agnostic.

A human-centered AI engine should automatically install, update, and resolve package dependency issues so that the entire engine can be deployed in different on-prem, cloud or other hosting environment. Software package dependency is a key concept for a human-center AI engine to minimize human interventions, and it can be as small as checking the dependency between two single source file or as big as deploying dozens or hundreds of software packages in the same environment.

A human-center AI engine should also self-learn to fully leverage the computing resource of the environment. The efficiency of the AI engine highly depends on how platform agnostic it can be

to adjust according to available computing source and hardware configuration of the environment it is up and running. For example, this might include what kind of models to train, how many models to train, what is the complexity level of parameter tuning, and so for. In an environment with limited resource, complicated models and parameter tuning should be avoid guaranteeing the minimum latency required by the BI frontend. In an environment with advanced computing resource such as GPU or TPU, the AI engine should automatically include more complicated models such as deep neural networks.

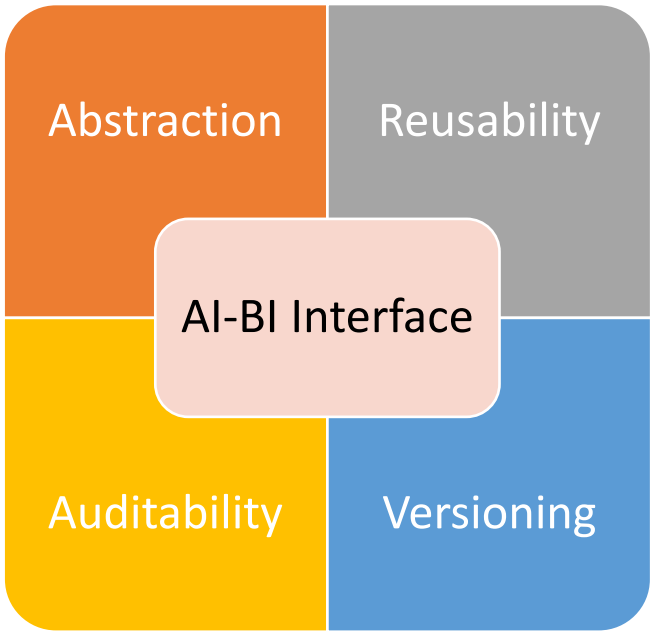
Rule #3 Model Recommendation

A human-centered AI engine is expected to minimize the amount of work for the end user to select what models to run and what models to be retired. Besides being frontend agnostic and platform agnostic, it is crucial for a self-learnable AI engine to recommend the right set of machine learning and deep learning for a specific input problem from the BI engine.

A key prerequisite of model recommendation is the ability to memorize and classify the input problems from the BI engine. This allows the engine to maintain a list of candidate analytic models that achieve good accuracy for a specific type of input problem and update the candidate list over time to include new models and retire old models. In many applications it is challenging to maintain a list of good models for a specific input data set. Therefore, the engine should be able to search and identify similar problems solved by the engine in the past and try to identify models that might work for similar problem to speed up the model selection process.

Beside keeping a model zoo that can quickly recommend a subset of models to execute instead of running all the models, a self-learned AI engine should be able to automatically update the champion model for the same input problem based on whenever the stability of the data changes dramatically. Most of the machine learning or deep learning models yield high accuracy when the scoring data set is consistent with the training data set. However, this might change when the scoring data set starts to include new columns or adding new levels to the existing columns that makes the current champion model no longer produce very accurate predictions or classifications. In this case, the AI engine should start to include the new observations into the training data set, retrain all the models, and select a new champion mode.

The AI-BI Interface by Scott Rigney



Because artificial intelligence encompasses a broad collection of techniques for learning from data, its potential to impact many domains is clear. Unfortunately, business intelligence (BI) has yet to fully harness the power of artificial intelligence and expose it to users at scale. This is partially explained by the fact that AI is an evolving field and its use generally requires expertise in frameworks, algorithms, and techniques that are unfamiliar to BI practitioners. As a result, AI has yet to converge on uniform standards that are pervasive in BI. To achieve the end-goal of self-learning BI systems, the industry should define a uniform set of standards for model abstraction, reusability, auditability, and versioning.

Abstraction

BI users are familiar with a drag-and-drop paradigm whereby their actions within a dashboarding application are translated into SQL. A sophisticated BI platform can customize its SQL generation patterns to query from hundreds of databases platforms and database versions. This is no small feat. BI's value proposition is that it simplifies the process of querying data so completely that even advanced users don't need to be proficient in SQL.

What would be the impact if BI applications could provide a similar level of abstraction to machine learning? For example, if a user could lasso a collection of data and guide the system towards identifying relevant patterns. This extends beyond enumerating the use cases and mapping them to preferred models. As another example, when training a model, a novice user may be unaware that walk-forward optimization is a more appropriate cross-validation technique than k-fold when working with time-series data. In both cases the system needs to adapt to user inputs and eliminate complex sub-steps. This effort to simplify the integration is needed in order to lower the barriers to adoption. The key point is that the BI system and its designers have an obligation to abstract away the complexity and build towards a model where drag-and-drop actions are automatically mapped to relevant algorithms.

Reusability

The concept of reusability is common in BI: it provides economies of scale. For example, a single filter on a business dimension ("Year" = 2020) can be used in multiple dashboards and applied to multiple datasets. What is the equivalent reusability construct for machine learning models

used in BI? Should the coefficients of a model that were found by training on the full dataset be accessible to a user who does not have access to the full dataset?

One approach is to think about the training pipeline in terms of reusability. In this example, a new regression model could be re-run against a subset of the data but re-use the training and optimization techniques used by its predecessor. A potential benefit to the end-user is that more relevant, nuanced patterns may be observable in a subset of the dataset than in the full dataset. While this premise runs counter to the general goal of model generalizability, the impact to the end-user may make it worthwhile. In terms of reusability, the result can be one model: one problem definition but containing coefficients (in the case of a regression model) that have a one-to-many relationship with end-users. The takeaway is that some adaptation to key concepts in machine learning, such as relaxing the requirement of generalizability, may increase relevance in a self-service BI context.

Auditability

Dashboards produced by BI systems are often subject to regulatory oversight. As a result, BI systems need to support calculation auditability. This requirement is complicated by the nature of machine learning: two different algorithms applied to the same dataset and given the same task may produce different results. Furthermore, some algorithms produce interpretable coefficients (regression models) and others do not (tree-based models). The result is that BI systems can't guarantee auditability. To circumvent this, BI systems can pursue a path towards transparency by leveraging model interpretability frameworks. LIME, short for Local Interpretable Model-Agnostic Explanations, is one such framework that seeks to describe black-box model behavior. In it, model explainability is achieved by repeatedly perturbing the inputs to the model and analyzing the resulting predictions. Although imperfect due to the nature of randomness in the sampling process, LIME and its variants offer a strategy for the auditability requirements that are common in BI.

Versioning

Similar to the requirements around calculation auditability, in BI, the logic used to produce calculations can also be change-managed if the BI system makes use of version control logic. A similar approach is used by data science teams who have solutions deployed in production; they're likely to use some variation of "model ops" or "ML ops" or "champion/challenger" techniques to control the model that is actually being used to make decisions. In such contexts, the rationale for promoting or demoting one model over another may be based on objective metrics or subjective business criteria. Because business data is messy and volatile (changes with time), BI stakeholders need tooling that would allow them to prevent a model from retraining, promote or demote a model, or restore a previous version of a model. This is an important consideration for real-world deployment of AI in BI, especially in industries that are heavily regulated.

Discussed here were a set of principles for technical integration that should be considered – by BI vendors and practitioners alike – in order to achieve more pervasive, yet practical everyday

use of AI in BI contexts. By model abstraction, we suggest that the BI system and its developers have a duty to abstract away the algorithmic decisions that would otherwise be made by a trained data scientist; the analogy of SQL generation and its impact on BI was given. Reusability defined a conceptual model for how to think about the training process in a way that provides a one-to-many relationship between the model and its BI end-users. The auditability requirement stipulates that models, no matter how sophisticated, need to be interpretable by business users and other stakeholders. Finally, with version control, BI users should be able to define the criteria by which the models used in the system are change-managed. It is hoped that the adoption of operating principles such as these represent a bedrock set of foundations needed to achieve technical integration between AI and BI. The result is a self-learning system that could usher in a new era of sophisticated, AI-powered analysis that is accessible more users and with less friction.

Enabling Humans to Visualize the magic of AI by Dalton Ruer

Figure 5 Creator: Blue Planet Studio | Credit: Getty Images/iStockphoto

Imagine if you will be in a world where human end users and Automated Machine Learning partner up. Alas, the majority of business users wouldn't know a Random Forest from a real forest. They don't seek to be data scientists, and for most Machine Learning is simply "magic." While fun to watch in person, magic isn't really trusted, and end users aren't going to take action on something they don't trust.

The insights and power of the machine, combined with the innate business knowledge and intuition of the business user, is too potent to ignore. Thus, the goal is to help them bridge that trust gap via:

- Education right in their Business Intelligence framework where they are already working.
- Allowing them to provide their input to the Artificial Intelligence to ensure it learns about the business and not just about the 0's and 1's about the business.

To that end we have derived these **4 rules any BI Vendor should follow if they want to allow their product to enable more end users to consume more Automated ML insights**.

Rule #1: Protect end users from harm

Like with many things, business users aren't always right in their assertion of what it is that they want. You likely have countless years of experience sitting in meeting, gathering requirements and watching users change their requests over and over and over. You are probably all too familiar with the fact that sometimes the very things that they ask for will cause them more harm, than not doing anything. But for sake of a good story bear with me.

For instance, a sales manager looking at current sales trends might say "I just want to see a forecast of how much we will sell next quarter." While your Business Intelligence tool can certainly provide some super slick interface that allows them to ask, the right question isn't "Can we forecast in a super intuitive way for the end user?" The right questions are "How

reliable will a forecast from the two quarters of their data be?" and "Will generating that forecast invalidate or reinforce their trust in the technology they can't see?"

Each Business Intelligence vendor will choose a different interface for the end users to take advantage of things like Time Based Forecasting. The sales manager will likely never care to know which "model" has been selected under the covers. But when the data suggests that they need it, the sales manager has to be informed by the Business Intelligence of how forecasting works and how the confidence of the results goes up the more data that can be provided. Remember the goal of Business Intelligence, regardless of the vendor, is for the insights to be actionable. Providing automated machine learning insights is the easy part, the tricky part is not presenting insights based on limited data that will erode the end users trust in the technology or cause them to take horrible incorrect actions.

Rule #1 for Business Intelligence vendors is the equivalent of the first rule of the Hippocratic Oath in the medical profession. Protect end users from harm.

Rule #2: Ask for Input

Imagine a marketing manager who has a known budget and can run several different types of campaigns. Should he send mailers at a cost of X, make personal calls or encourage face to face events? Each has a known cost. So, he asks for a prediction of what to do for which customers.

Easy right. Behind the scenes we could run 57 different models, choose the one with the highest confidence score and bingo he has his results on screen. The marketing manager doesn't need to know any of the 57 model's names. But they likely do need to be prompted for one key piece of information … do they care about negative prediction? Sending an email to a customer who doesn't respond has minimal cost. But what is the cost of staging an event for face-to-face involvement and our model having a 98% accuracy but a 5% negative rate. Would they prefer that over a 97.7% accuracy and a 0% negative rate? That's where it gets interesting.

Like all business users, the marketing manager who wants to take advantage of the power of Machine Learning on the fly, doesn't care one iota about the fact that 57 models were run. But they are the only ones that can provide that very key piece of cognitive intelligence that they have which the models don't.

While any Business Intelligence solution can generate insights from raw data without educating or prompting end users. The data is only half of the solution. The success of human centered Artificial Intelligence depends upon the BI vendors following this rule. Educate the end users enough about what is happening, ask the end users for their input.

Rule #3: Confidence Scores instill Confidence

Another type of education that must be provided is regarding the data size and its potential relationship to the predication accuracy. A data scientist who is asked to analyze employee churn is going to know that the prediction is going to be much more accurate if they are asked

by a company providing them 20,000 historical employee records, than if they are asked by a company providing only 15 historical records.

However, the Human Resource managers at both companies seeing a Business Information screen with a button to run churn are both going to want to press the button. After all it is "magic." They won't know, until told, why the amount of data used matters. Likewise, they won't understand what a z-score, t-score or p-value is. That's where the BI vendors need to provide the education in some way rather than just automatically choosing to run and report whatever model had the highest score. Business users will need to know what the "score" is and educated whether it should be trusted in a way that fits their role. Meaning business users won't understand why there is such a big deal between a .2 and a .7.

Put yourself in their shoes and imagine clicking that pretty icon to run churn while analyzing data in the company's Business Intelligence tool and then seeing:

> "{insert your name here} the data you provided in this application was run through a series of Machine Learning models and it has predicted that John Doe will remain with the organization. Because the magic system had so much data to work we are super-duper confident that this prediction is accurate and can be trusted."

Or seeing

> "{insert your name here} the data you provided in this application was run through a series of Machine Learning models and it has predicted that John Doe will remain with the organization. Because the magic system had so little data to work in the application you are using as a result of the filters you have in place, we suggest you take that prediction with a grain of salt."

All kidding aside, the point is that as a result of the size of the data being utilized by the business end user the p-z-t values could be all over the place and the end user needs to know. Regardless of a high or low confidence score, being educated will instill trust and confidence in the business users.

Rule #4: If it looks too good to be true, it probably is

Data science is a lot like an online dating application. The first thing a data scientist will do when considering a relationship with the data is run a data profile. Because they will always want to know if the data is trustworthy, or if it looks too good to be true.

Imagine a telecommunication company wanting to investigate customer churn. A data scientist would obviously need to ask, "Which field represent that your historical customers have actually churned?" When they run a data profile, they might report that there is a 100% correlation between Field A and the field specified as the target. Yeah, our p-z-t-a-b and c scores are going to be off the charts when we use that model. So, while the manager will be

ecstatic if we follow rule #3 and report the scoring in some manner, it might not be trustworthy because it's "too good."

Thus, Business Intelligence vendors must also do some level of data profiling and educated the business users before blindly running Machine Learning algorithms. When something in the "profile" seems suspect, business users need to be informed and allow them to choose the path.

"Oh silly me that is just the text version of the target field before we translated it to a numeric flag. I shouldn't have included it in the list of variables, please ignore that field." Or they might say "That is fantastic news we found a 100% accurate correlation variable. Please run the predictions using it and I am certainly going to get my bonus."

The opposite scenario also needs to be handled via profiling where it's clearly a data problem causing a really low score. "Are you kidding me? 92% of the data values you want me to use for the predictions are NULL. This isn't magic, I can't fill in the blanks for you and offer good predictions."

All kidding aside, business users need to be educated by the business intelligence tools regarding the data's trustworthiness. If they aren't going to want to trust the results, there is no sense even beginning the relationship.

If BI vendors can follow these rules, we can dramatically magnify the power of BI or AI alone by creating a partnership between business users and automated machine learning. Naturally each vendor will apply its own secret sauce, if you will, as to how to implement these 4 rules. The goal isn't for all to have a common look or feel. Just to ensure that:

- End users are protecting from harm
- End users are asked for input when higher model scores aren't the only thing that matters
- End users are told in nonscientific ways how likely predictions are to be accurate
- End users are prompted when need be if their data looks too good to be true, or so bad it needs to be ignored.

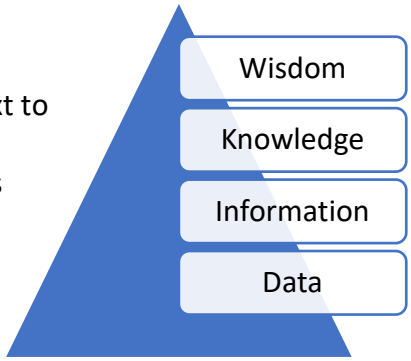
Data Quality by Sachin Sinha

In structured data world there is huge importance given to data quality. It is widely perceived that “If your data is bad, your machine learning or business intelligence tools are useless”. While bad data quality is not a new problem, its effect increases exponentially in the machine learning and ML driven BI world. We have already seen in public domain some biases these models can deliver because of bad quality data. In the enterprise world this can lead to catastrophic consequences if business starts taking decisions based on predictions from a model that was fed bad quality data to create.

However, there are instances where data is just not within our control. We have all run into instances where a sparsely populated data set is all you have. We can all scream about the quality of data but sometimes data is what is given to us and there is no way to change that. Now, even with that sparsely populated data, a subject matter expert human can take decisions that will be better than what any model will come up with using this sparsely populated data. The reason being the presence of human knowledge, experience and wisdom which is absent from anything that ML is building.

In the earlier sections we learned that machine teaching seeks to gain knowledge from people rather than extracting knowledge from data alone. Essentially, we can now use knowledge that was generated and internalized in the past to help with the decisions that machines are either going to take or help with in the future.

We all know about the Data-Information-Knowledge-Wisdom hierarchical pyramid. That is how humans learn. That is how we accumulate knowledge, insight and wisdom. By providing context to the data, we create information which results in knowledge and eventually wisdom. BI vendors have operated for decades in this pyramid. In fact, BI tools are primarily responsible for taking the corporate data and turning it into information by providing context and meaning to the data which then results into corporate knowledge and eventually wisdom to act.



Can BI tools expand their influence on the pyramid and now maybe even make use of the upper layers to solve the problem of data quality. Can we combine data and machine teaching using the knowledge and wisdom of a subject matter expert to build better model and as a result better analytics even in the absence of pristine data quality? Is it possible to finally address the issue of data quality without inserting synthetic data into the dataset to fix the data quality issues? The result of poor data quality is that it will require not only synthetic data but also labeling of the data in a way that can counter absence of data or poor data quality. If BI tools can leverage knowledge and wisdom from the subject matter experts, we can build models and create analytics, with fewer labels and less than pristine data, than traditional methods.

Just like the role of the teacher is to optimize the transfer of knowledge to the learning algorithm so it can generate a useful model, SMEs can play a central role in data collection and labeling. SMEs can filter data to select specific examples or look at the available example data and provide context based on their own intuition or biases. Similarly, given two features on a large unlabeled set, SMEs can conjecture that one is better than the other. However, what we are doing now is that finally leveraging the knowledge and wisdom layers of the pyramid to enrich and enhance the data layer, filter out the noise and provide a better context, not on the gut feeling, but based on the knowledge and wisdom that was created based on data from years or decades of experience.

So how can BI vendors enable that? BI vendors should try to leverage advances in machine teaching to bring some real-life benefits to the end users. First, they should try to incorporate data profiling algorithms to auto detect the quality of the data, not just from population of the dataset perspective, but also from the perspective of how accurate analytics modeled on this dataset will be. This will serve two purposes. It will inform the user of the quality of their dataset and it will also alert them that this is something that they will probably need the help of a SME.

Once the data profiling algorithms detect lower than threshold quality then BI tools should kick in the second step. For this step they should build workflows for SME interaction that will kick-in automatically when the algorithm detects data quality lower than a certain threshold. At this point, BI tool will offer its user a workflow that will teach the machine to counter poor data quality. As an example, this could be in the form of elimination of rows of data because it depicts seasonality that is not applicable to the analytics being generated.

Leveraging a SME and their wisdom to teach a machine and solve data quality problem is not the only way to bring wisdom back into the tool set that helped generate that wisdom in the first place. Another avenue that BI tools can leverage is data cataloging. Data cataloging is a way to create knowledge and wisdom about which datasets have better quality and which datasets are applicable in a particular context. BI vendors should work on better integration with the data cataloging tools. This should not be limited to the integration where an end-user can open a dataset from the catalog into their BI tool of choice. Ideal integration would truly leverage the knowledge and wisdom that data catalog helped create. It will provide, for example, end users, suggestions for right data sets, when it identifies that the dataset in use has quality issues and there are others in the catalog classified as better dataset for this use case.

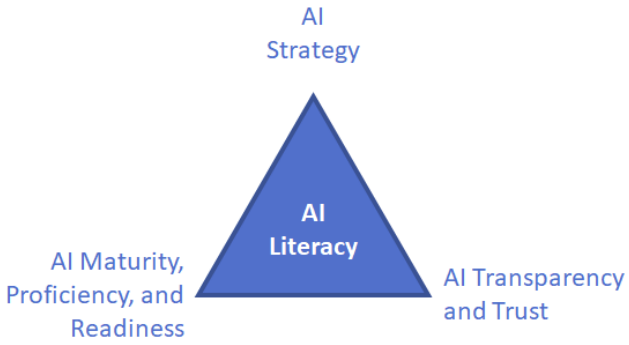
While machine teaching does appear to be something that can help immensely in the data quality area, the onus of making it available to the end users is on the BI tool vendors. Also, there are other ways to reach into the upper layers of the information pyramid and use that to help end users. As the end user tool of choice for analytics, business intelligence tools are the right medium to not only help them reach into the upper layers of the pyramid but also make it useful for them by using it to solve a very important question that remains a thorny issue, data quality.

AI Literacy by Gerard Valerio

We are in the midst of the 4th Industrial Revolution where the convergence of various technologies are redefining the world we live in. Among these technologies are analytics and artificial intelligence (AI for short). Analytics and AI are the engines behind data and digital transformation in the business world. Everything from searching using Google to recommendations and suggestions on online shopping and social media sites to automated call center response are being fueled by the use of analytics and AI. All of these are great examples of human-centered AI where AI is applied to the human experience and interaction with technology.

Here's a question to spark thought: What is AI Literacy and how do you know if you have achieved it? Good question. AI Literacy begins with asking ourselves what does great look like when AI is used successfully? What goes into successful application and use of AI?
Achieving AI literacy is both the journey and the destination. With respect to AI Literacy, here are 3 critical success factors to consider as a starting point for achieving AI literacy:

(1) AI Strategy
(2) AI Maturity, Proficiency, and Readiness
(3) AI Transparency and Trust



This list of critical success factors is by no means complete and so ask yourself what else would you add as a critical success factor to achieve AI literacy?

AI Strategy

The successful use of AI begins with strategy. What is it that you are intending to happen or occur through the use of AI? Who are the benefactors? Why do the benefactors or users need AI? The use of AI rises out of use cases backed by a value proposition for a product or service that either augments or automates a business process or a human experience.
Any AI application or use involves data inputs and data outputs. The data inputs can collect user information or input, user interaction data, sensor data, etc. This input data is then processed by an algorithm or AI code that is developed to produce an output such as data or information that feeds an action to be taken or a decision to be made whether automated or requiring human involvement.

Here are some considerations that go into framing AI strategy:

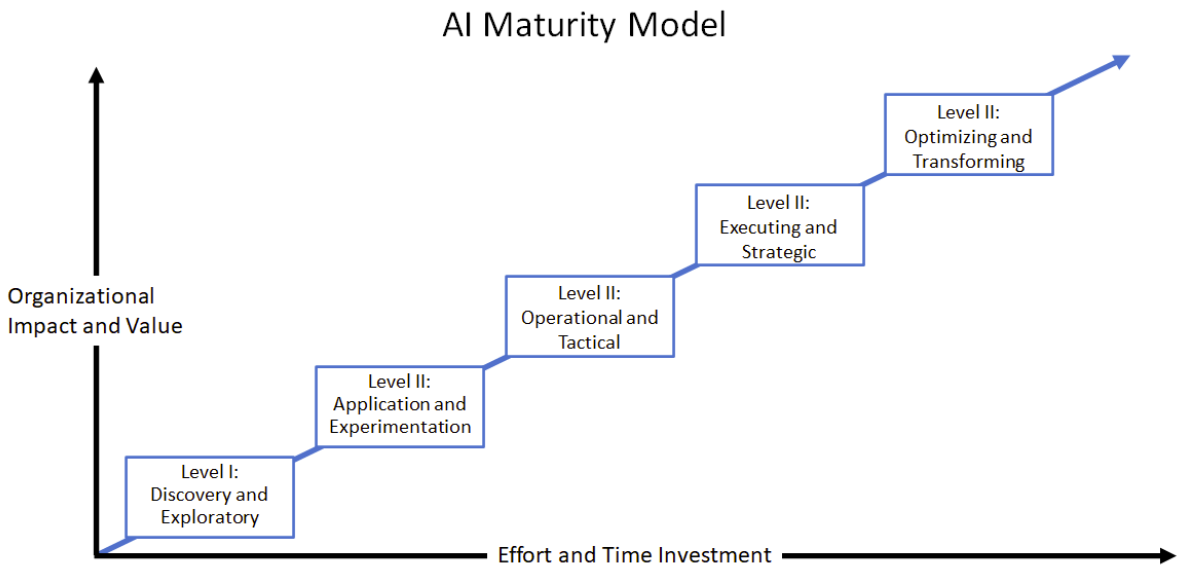
- Data Strategy – Data underpins AI and therefore the use of data as inputs and outputs to an AI algorithm needs to be well defined.
- Ethical and Legal Considerations – Does the intention of the use of AI or the outcome produced by an AI algorithm raise into question any ethical or legal considerations? AI can be used for both social good and social evil applications. Today, we live in a world that is

beginning to be plagued by disinformation through social media. While social media was designed to foster community and connection through family, friends, and followers, it is also being leveraged for nefarious purposes. The same AI used on social media sites such as Facebook and the like can also be manipulated to do just the opposite of what may have been intended – good information can benefit building community and connection while disinformation can breakdown community and connection.

- Human Factor – Central to the application and use of AI is the human experience. How are humans interacting with AI? Are there any requirements for a successful interaction? What is a failed interaction and why did it happen? How do we account for the variability of human action and behavior? Lastly, how is change integrated so that humans embrace AI comfortably?
- Technology – At the core of AI is the use of technology as a vehicle through which AI acts. Considerations such as choice of technology, skills and talent necessary to build AI, and deployment all matter in the successful application and use of AI.
- Use Cases – What is it that you are trying to make better through the use of AI? Is it a product or service and making them better, faster, smarter? Is it process automation intended to improve efficiency and speed? Whatever the case may be, this is what informs the design and development of an AI algorithm.

AI Maturity, Proficiency, and Readiness

Organizations undertaking AI will find that this is an iterative and learning process that unfolds through experimentation (trial-and-error and trial-and-success). Achieving AI literacy as an individual or organization is an evolving process. Just like with any new capability or skill, an organization will first have to traverse a maturity lifecycle. Think crawl, walk, run.



AI literacy as a means to drive successful use of AI comes from an awareness and maturing AI skills (competencies and proficiencies). Think of AI maturity in the context of the Capability Maturity Model (CMM) and the 5 levels of maturity.

In the context of AI, here are the 5 levels of maturity of AI as a capability:

- Level I: Characterized as discovery and exploratory, individuals and organizations are developing an awareness of what AI is and what is possible with AI.
- Level II: Characterized as application and experimentation, AI is being developed and tested to materialize a concept into existence and to learn from failure and success.
- Level III: Characterized as operational and tactical, AI is "productionized" for experience and process efficiency.
- Level IV: Characterized as executing and strategic, AI use is becoming pervasive and leading to broader change and innovation with respect to digital and data transformation. Critical mass of proficiency is achieved, and organizational use is beginning to scale.
- Level V: Characterized as optimizing or transforming, the use of AI is broad and enterprise or organizational wide.

AI Transparency and Trust

Trusting AI is perhaps one of the bigger challenges on the road to AI literacy. Why? Mostly because the application or use of AI is encountered as a black box experience to those who are knowingly or unknowingly the intended benefactors. We have all heard of the phrase "trust, but verify" when it comes to analytics. To establish AI trust, producers of AI must test often and test continuously (monitor) their implementation of AI in order to ensure intended outcome or result produced by the use application or use of AI. Trust is built and gained by the consumers of AI when it consistently and positively benefits their experience with AI.

Another way to contribute to AI Trust is to also think about how AI can be made more transparent and easier to understand. AI transparency has been associated or equated to the concept of explainable AI (XAI). When considering the use of AI and factoring in the idea of transparency or explaining the AI algorithm in the context of human understanding, we build AI trust as we work to demystify the black-box experience and nature of AI algorithms.

Conclusion

The road to AI literacy is a complex journey with challenges. Having a roadmap helps define the journey in the context of creating intentional outcomes through the use of AI for a business purpose. This roadmap is AI Strategy. The road to AI literacy is a transformation process with different stages of organizational maturity in the application and use of AI. Knowing where an organization is in its AI capability maturity helps define the near-term focus and what is necessary to move to the next capability maturity level. The road to AI literacy requires developing transparency and trust in the application and use of AI. Without a commitment or effort towards AI transparency and trust, AI literacy is not possible without overcoming the biases and fear of what people suppose AI is all about (AI is here to replace humans resulting in loss of income, loss of jobs, loss of personal prosperity, etc.).

Authors by last name alphabetically

Cupid Chan

Cupid Chan is a seasoned professional who is well-established in the industry. His journey started out as one of the key players in building a world-class BI platform. He has been a consultant for years providing solutions to various Fortune 500 companies as well as the Public Sector. He is the Lead Architect of a contract in a government agency leading a BI and analytics program on top of both Big Data and traditional DB platforms. He is the Board of Directors, Technical Steering Committee (TSC) and the Chairperson of BI & AI Project in Linux Foundation ODPi.

Xiangxiang Meng

Xiangxiang Meng is a Staff Scientist in the Data Science Technologies department at SAS. Xiangxiang received his PhD and MS from the University of Cincinnati. The current focus of his work is on the algorithm development and platform design for machine learning and business intelligence software, including SAS Visual Statistics and SAS In-Memory Statistics on Hadoop. His research interests include decision trees and tree ensemble models, Bayesian networks and Naive Bayes, recommendation systems, and parallelization of machine learning algorithms on distributed data.

Scott Rigney

Scott works as a Principal Product Manager at MicroStrategy. He manages several products including data science integrations, APIs, SDKs, and is the creator of the "mstrio" package for Python and R. Before MicroStrategy, he worked in data science and developed machine learning systems for predicting application outages, process optimization, and IT system dependency discovery using network graph models. Scott holds a master's degree in data science from Northwestern University and a bachelor's of science in Finance from Virginia Tech.

Dalton Ruer

Dalton Ruer is a Data Scientist Storyteller and Analytics Evangelist. He is a seasoned author, speaker, blogger and YouTube video creator who is best known for dynamically sharing inconvenient truths and observations in a humorous manner. The passion which Dalton shares thru all mediums moves and motivates others to action.

Sachin Sinha

Sachin Sinha is a Director and Technology Strategist at Microsoft. After graduating from the University of Maryland, he continued his information management research as a Data Engineer and Architect. Throughout his career, Sachin designed systems that helped his customers make decisions based on data. During this time he helped startups in the healthcare space get off the ground by building a business on data and mature from seed to series funding. Sachin also helped several organizations in public sector achieve their mission by enabling them for decisions based on data. Now at Microsoft, as a Technology Strategist, he helps customers with digital transformation. Sachin takes pride in engaging with public sector customers each day to help them achieve more for their organization's mission. He currently lives in Fairfax, VA, with his wife and two sons, and remains a fervent supporter of Terps and Ravens.

Gerard Valerio

With more than 7-years at Tableau (now Salesforce), Gerard Valerio is a director leading a team of solution engineers in evangelizing Tableau and Salesforce to the U.S. Federal Government. He has built a career on data spanning first generation data warehouses in Oracle, Informix, and Sybase to implementing and selling business intelligence tools like SAP Business Objects, IBM Cognos, and MicroStrategy. Mr. Valerio also worked in the data integration space as a customer and employee of Informatica. His Big Data experience spans working with Terabyte to Petabyte-sized data volumes staged on in-memory columnar databases like Vertica to structured/unstructured data residing in Hadoop-based data lakes.