How government agencies can get started on their AI journey

By Sukumar Iyer

Founder and CEO, Brillient Corp.

AI awakens

John McCarthy and Marvin Minsky coined the term "Artificial Intelligence" in a grant proposal for a workshop they were seeking to conduct at Dartmouth college in 1956. As Kenneth Cukier[1] recounts, they came up with this term to keep someone out of that workshop. That someone was Norbert Wiener, a child prodigy who got his Ph.D. from Harvard at the age of 18. Wiener had coined the term "Cybernetics" in 1948 as "the scientific study of control and communication in the animal and the machine". Apparently, Wiener was an insufferable "know it all" who would have monopolized that workshop. To keep Wiener out of the workshop McCarthy and Minsky created the term Artificial Intelligence (AI) instead of cybernetics.

As Shakespeare wrote in Romeo and Juliet "*A rose by any other name would smell as sweet*". So, what is in a name? If "statistical learning methods" or cybernetics had been the term that the scientific community had congealed on, one wonders if our imaginations and creativity would have sent us rocketing up the hype curve that AI is going through today. Would we have seen the likes of the all-knowing Hal in 2001 space odyssey or the robots in the Alien and Terminator series of movies from Hollywood?

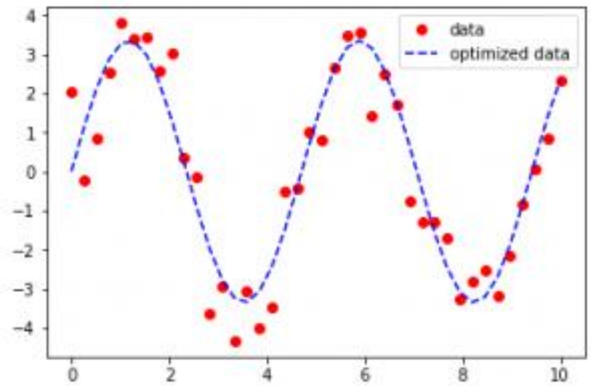
AI is everywhere in the news on TV, social media and websites. As with most technologies going through a hype cycle, there is a lot of noise in the marketplace and it is hard to zero in on exactly what AI is and how government agencies could use it. This article seeks to dispel some of the myths around AI, cut through the noise and zero in on opportunities.

What is AI?

Alan Turing the famous British mathematician is widely considered to be the father of modern-day computer science and artificial intelligence. During World War II, he was a pivotal member of the British code-breaking team at Bletchley Park that broke the German cipher codes allowing the allies to read enemy message traffic, sometimes even before Hitler did. Under the code name "Ultra" this intelligence played a major role in winning the war. In 1950, he developed the Turing test which is a test of a machine's ability to exhibit intelligent behavior equivalent to, or indistinguishable from, that of a human. Passing the Turing test is the holy grail for all AI technologists.

As we saw, AI is quite old, from about half a century ago. Attending AI conferences in the last three years, it has become clear to me that scientists and researchers are still trying to define what AI is. In simple terms, today, AI is a set of statistical and mathematical algorithms that use vast amounts of historical data to learn and make predictions into the future. In excel you can create charts where you fit a curve among the various points of data to get a graphical feel for what the data is telling you.

[1] Ready for Robots? Book review by Kennet Cukier, Foreign Affairs, August 2019



The dashed curve in the picture is a curve fitted to discrete, scattered points of data.[2]

There are many mathematical methods for curve fitting including linear or polynomial curve fit and regression analysis, for example. Once you have the fit you can make some predictions on where a future dot of data might fall by extrapolation. An example could be the sales figures for a department store. A curve fit of the past sales trends can be used to predict future sales including seasonality and other factors that went into the historical data analysis.

Similarly, the statistical and math algorithms that constitute our AI techniques today can use past (training) data to build a "model" that can make future predictions. I often like to dumb down AI as the process of teaching a dog new tricks. After numerous rounds of repetition, the dog learns the trick and can respond to your gesture or call. And, yes sometimes the dog does something unexpected as AI can too.

The pundits have defined three levels of AI:

1. Narrow AI – model solves specific problems and cant be used for other problems
2. General AI – models that can solve many problems
3. Superior AI – smarter than the human brain

This article focuses on narrow AI. The other two areas are currently topics of research.

Other synonymous terms to AI include machine learning (using computers, the machines to learn the historical data) and cognitive intelligence, cognition being defined as the process of acquiring knowledge and understanding through thought, experience, and the senses. We will simply use the term AI or bots in this discussion.

AI has been amongst us for more than a decade embedded inside many things we use. Some popular use case examples include the amazingly accurate spam filters in Gmail, Microsoft Outlook's segregation of emails into clutter, focused and other streams, Google assistant, Siri, Alexa, Netflix, Amazon, and Tesla autopilot (the latest versions use machine learning to sense and turn on wipers too). Apart from regression and statistical methods, neural networks are a mathematical approach used for "deep learning" which seeks to mimic the working of the human brain. Neural nets parcel information into various neural data "nodes", apply different weights to the data in each node and use multiple layers of these nodes and "backpropagation of error" to vary the weights and achieve the curve fit. Almost 30 years ago in grad school, I worked a summer internship at a company involved in the manufacture of roller bearings (which allow shafts to rotate in various machines including vehicles). I used an early neural nets program running on a PC to predict the influence of various input factors (such as type of steel, hardness,

[2] https://www.geeksforgeeks.org/scipy-curve-fitting/

temperature, etc.) on the quality of roller bearings being produced by that company. The resulting model identified the mathematical equations, weight multipliers for each factor, etc. that impacted a certain quality metric. Neural nets are thus quite an old technique going through a second life now. These neural nets (or NNs) have subtypes such as Convolutional Neural Nets (CNNs) and Recurrent Neural Nets (RNNs) just to name a few that are used by a model to “learn” the historical data.

Let’s review a few other terms. Supervised learning algorithms build a mathematical model of a set of data that contains both the inputs and the desired outputs. The input data is known as training data and consists of a set of training examples. Once the model fits the curve between input and output data it can now make predictions on future data. A commonly used example is using thousands of labeled photographs of cats and dogs. The model is supervised to identify which one is a cat or a dog. When given a new picture that it has not encountered before, it can predict with a degree of confidence if the picture is of a dog, cat or not. These are also known as “classification” problems.

Unsupervised learning algorithms take a set of data that contains only inputs (no training data, no training), and find structure and patterns in the data, like grouping or clustering of data points. The algorithms, therefore, learn from data that has not been labeled, classified or categorized. Instead of responding to feedback, unsupervised learning algorithms identify commonalities in the data and react based on the presence or absence of such commonalities in each new piece of data[3].

Why AI now?

Since its inception, Artificial Intelligence has gone through an “AI winter” with multiple setbacks (including inflated claims leading to loss of credibility) before once again catching fire at the beginning of this millennium. This new dawn of AI was facilitated, among other things, by the plentiful availability of data, computing power and the advent of the cloud.

The proliferation of IT systems, the internet, and social media have led to exponential growth in the availability of data. Where once companies and agencies knew little about their customers or citizens, now they are inundated with data, the “big data” problem. Combined with the stupendous increase in computing power and the ubiquitous cloud, AI techniques have started delivering value. I often like to say that most AI now and in the future will be embedded technology like the “intel inside” ads that showed how intel chips inside computers made them run. Often, we will not see or explicitly buy AI but a solution or product with AI embedded inside. With the coming of high bandwidth, almost ubiquitous 5G networks and the Internet of Things (IoT), the amount of data will explode exponentially making more data available for AI.

The data challenge for agencies

As discussed earlier, data is the foundation of AI today. In order to get good fidelity results, we need good quality data for training AI models. Information Technology (IT) systems were a boon for companies and agencies alike, moving them from a world of paper and human-intensive tasks to more automated systems, self-serve and in general a higher level of customer and citizen services. IT systems in most agencies have grown and evolved rapidly and often in independent directions. What this means is that most agencies do not have standardized data taxonomy (what something is called in each system) for even simple fields like a person’s name or addresses. I recently moderated a government panel on “Getting your data ready for AI” at the AI World Government conference in Washington DC. We learned that Health and Human

[3] https://en.wikipedia.org/wiki/Machine_learning

Services (HHS) procurement systems have 11 different terms for a contract "award". On a contract at the US Dept. of Energy, we discovered that different systems (procurement, finance, property management) had different terminology for how they called the same computer. Variations included Laptop, Computer, Server, Blade, Automatic Data Processing to rattle off a few. On our IRS contract, we learned that in one system a field called State did not refer to what we might think – the 50 states in the country, but the state of something else. This is not unique to government agencies but is common across large private sector companies as well. One of the main contributing factors is that given a system mandate, IT teams would get together and build a system for a specific purpose without worrying about how it interacted with other systems or fit into the overall goals of the organization's mission. Thus began the creation of system and data stovepipes.

To get over the stove-piped data problem, a common solution used to be the construction of big "data warehouses" which brought together information from multiple systems into one database system. Unfortunately, when data is moved, it can get mangled, corrupted, changed or irrevocably transformed in ways that are not intended making that data quite useless. Every time data is copied from one place to another, the chances of the two sets of data going out of sync are very high. According to some data gurus, 60% to 85% of data warehouse and big data projects are failures and costly white elephants[4].

Across both transactional systems as well as data warehouses there are numerous data quality issues including duplication of data, incorrect data, corrupted data to name a few. Data can be created bad, loaded bad or corrupted during transit or transformation. As Michael Conlin, the chief data officer of the Dept. of Defense noted at the ACT-IAC AI IA Forum 2018 "At DoD, we have tons of data, but most of it is crap". When the input data is flawed, it results in the GIGO problem or Garbage In, Garbage Out.

Another serious problem agencies face is fragmentation of data across the enterprise and data living in silos that are often inaccessible. As discussed, in most agencies systems were developed in independent directions, often focused on running the mission function of one part of the organization. For example, a taxpayer may be handled by multiple silos'd systems that deal with wages and income information, audits, enforcement or criminal investigation. It is very hard to get a 360-degree view of the main entity in question the taxpayer who exists across these systems. On top of that, system owners zealously guard data in their systems and getting access is often tedious and challenging. A deputy CIO once told me that cybersecurity was his best friend and his worst enemy. Security protocols in place often prevent timely access to the right data. One must first access data across these silos before transforming the naming, format and type conventions or combining data for training AI models or enabling automation on retrieving transactional data. The more quality information we have around citizens, taxpayers and companies, the better AI models we can build on them.

These are serious problems. To get outcomes we can rely on, we need at least some amount of good data with reasonable quality. Does this mean that no AI is possible before first cleaning, standardizing and combining all the data we have? All is not doomed, there are opportunities to get started.

Opportunity themes

Let's first start by discussing the thematic elements of how AI can help government agencies before discussing what those opportunities might be.

[4] https://www.techrepublic.com/article/85-of-big-data-projects-fail-but-your-developers-can-help-yours-succeed/

1. AI to augment, not replace

AI techniques are not a panacea for all your problems. They are one more tool (like IT and the web were) that is available to help. Think of AI as a means to augment and boost what you currently do rather than replacing what you do. Like the "intel inside" slogan, you can embed AI to make your systems and processes more "intelligent" and effective.

2. AI to (semi) automate

The biggest opportunity for AI techniques is increasing the levels of automation. Until Terminator arrives, AI is going to be a dumb mutt for a long time. As such it is well suited to automate tasks feeing up humans to spend their time doing more valuable things. This is already happening with the likes of Google Assistant which regularly gives you inputs on traffic that might impact your drive home, how long it takes to get to work or new restaurants in your area that you may like (based on your visits and reviews).

3. Intelligent Change

The most significant challenge to implementing AI in government will be the workforce. In any organization (not just government) change management among the people has been the biggest challenge even in the roll-out of Information Technology. At Brillient, we see this even today in rolling new interfaces to systems, people prefer what they are used to. Top-down mandates may face some success and then run into the reef banks of failure. The imperative is for intelligent change. Start with some pilots to show employees how their lives can be made easier, how the agency's ability to execute the mission is improved and how it makes citizen's lives better. Gaining champions in a crawl, walk, run progression is key to success.

4. Commoditization of AI

Even as it goes through the hype curve AI tools are getting commoditized. Google, Amazon, Microsoft and open-source groups have already created tools you can download and start building, deploying and running AI models. Without knowing a whole lot about AI algorithms or python programming a reasonably technical person can implement a no-code AI model with some success. Many pre-built models already come trained to recognize faces or objects using a Convolutional Neural Network (CNN) for instance. This commoditization is both an advantage as well as a disadvantage. For somewhat straightforward problems like image recognition (as explained later) these models offer a quick path to a solution. Applying models to perform data analytics in the hands of an amateur can lead to disastrous results. AI tools are like the Integrated Development Environments (IDEs) of the IT programming age. They are better used in the hands of AI and data-savvy technical people.

Immediate AI opportunities for agencies

Congress and the white house have strongly supported a government focus on AI and maintaining the US as a leader in the field[5]. A manifestation of this focus is that all agencies have to nominate a Chief Data Officer (CDO). That alone is not going to be sufficient. The CDO must also be provided the power to make change happen. The CDO must be able to establish and execute a data governance program which includes establishing a common data taxonomy for the agency, developing data dictionaries for systems, implementing changes to data in systems while maintaining backward compatibility with older data, establishing best practices and being

[5] https://www.whitehouse.gov/presidential-actions/executive-order-maintaining-american-leadership-artificial-intelligence/

the primary sponsor of the recognition that data is an asset just like budgets, money, people and equipment.

Where and with what data can agencies get a quick start with AI? Here are some use case scenarios to consider.

1. Imaging, surveillance, tracking

There are numerous use cases across DoD, DHS, DOJ, USDA and health agencies where copious amounts of imaging data are collected. This includes camera feeds, aerial and satellite remote sensing, tracking people and equipment as well as medical imaging. This data suffers less from the data challenges discussed earlier and can be an excellent source of labeled training data to help AI models identify people, places and things of interest. AI-enabled imaging programs already help insurance companies use satellite imagery to resolve a claim for roof damage without having to send out a claims adjuster thus saving a lot of money while making the claim process faster. Security and law enforcement agencies already use license plate readers to identify vehicles of interest. Pilot projects are underway at some of our airports where facial recognition is used in lieu of a boarding pass for passengers to emplane. At a conference, the CIO of MWAA told me that they used a regular Apple iPad at Washington Dulles airport, tweaked it's built-in facial recognition model (using 30,000 data points on a facial image) and achieved a 99% accuracy rate for a boarding pilot.

While most agencies think of big data as numbers and text, I was surprised to learn that the majority of big data in the health sector is made up of images. As the products of diagnostics tests, these images can be correlated with the physician's diagnosis (labels for creating training data) to train imaging models on what constitutes a cancerous tumor, blocked artery or skin disease. AI models can now predict certain types of skin cancer at lower error rates than dermatologists.

Images and videos thus offer a cleaner data set with defined data formats for AI models. Video offers an additional big data challenge of velocity (speed of data) that can be resolved in algorithmic ways and using the muscle of cloud power.

2. Automation of business processes

Robotic Process Automation (RPA) mimics human "hand work" on a computer to automate repetitive tasks based on business rules. RPA is gaining traction across government with initial gains in procurement and financial processes. See this article on RPA across government. There are tremendous opportunities to graduate RPA to "intelligent automation" using AI methods before, after or even in between RPA tasks. At Brillient, we plan to use Cognitive Machine Reading (CMR) - which is AI-based pattern recognition of desired text in an unstructured PDF document or even a scanned document – ahead of doing RPA with the derived data. This is an example of using intelligent solutions to read documents and extract fields like invoice numbers, dollar amounts, product line items, etc.

3. Automation of customer / citizen interaction

Most agencies use Interactive Voice Response (IVR) systems to handle phone calls from citizens and businesses. Many agencies facilitate the handling of information requests and transactions via their websites. These systems collect call and weblog data in a well-structured format suitable for analysis. This includes the calling number, the choices the customer made (hence the reasons for calling), duration, pages visited, information searched/clicked, transactions executed, etc. For phone calls, customer service reps usually have some kind of CRM or case management system that tracks the reason and resolution of each call. This data is captured in well-defined formats and structure which can be used as training data for an AI model. We are already seeing the use of robot IVRs where they start with something like "how can I help you today" instead of the old menu of "press 1 for …….. and press 2 for ……..". These systems use a combination of Natural Language Processing (NLP), context and in some cases a dip into the calling customer's data. Ideally, the AI assistant knows everything about you since you called from your cell number on file and doesn't have to ask for your address or other identifiers and can get you going towards call resolution. At a minimum, the robot IVR can connect you to the right human rep who can solve your problem.

To be more useful, an AI assistant whether on the phone or on a website can better assist you in not just understanding what you want to do but also fulfilling the request. In order to automatically handle information requests (how do I do something, what form must I fill, etc), agencies must build a good Frequently Asked Questions (FAQ) knowledge base. The company that provides my internet service has a pretty good robot technical assistant which has solved my problems almost all the time. I assume they use both an FAQ database as well as clearly defined Standard Operating Procedures (SOPs) on what the bot needs to do. I am sure the company has saved millions of dollars pleasing their shareholders.

To automate a transaction (I need to file something, what is the status of my refund, I want to register a vehicle, etc), apart from an FAQ database the AI assistant will need access to mission systems and customer data. We now run into the data challenges detailed earlier including stove-piped data, PII, authentication, security concerns, federal laws, etc. There are numerous ways to design solutions for these problems while they may take longer to implement. That's grist for another article.

4. <u>Imperfect AI</u>

Jack Welch, the former CEO of GE famously said: "Deal with the world as it is, not how you'd like it to be". While agencies have the data challenges enumerated earlier, what can we do with what we have? What can we do despite the taxonomy and data silo issues? There are opportunities for very narrow AI focused on the narrow slivers of data an agency may have. Some that come to mind include:

a. Financial systems – analytics and forecasts of spending, budget trends, changes in what an agency buys over time, etc.
b. Asset management systems – changes in what the agency buys, lifecycle and replacement costs/timelines
c. Procurement data – analytics and forecasts on what we buy, who from, what we pay that may guide procurement policies
d. HR data – to predict changes in workforce, demographics, deciles, qualifications

These systems may suffer from some of the data challenges. They will also be dependent on upstream systems and data issues. However, they may be a starting point for some imperfect AI as long as we keep in mind that it is imperfect.

A longtime colleague and friend likes to remind me "don't let the perfect be the enemy of the good". Sometimes an AI solution as enumerated above might be good enough, at least for now and to get started.

Things to watch out for

1. <u>Data bias leads to biased AI models</u>

The fallibility of AI is quick to catch the attention of news media. There have been numerous examples where AI models tend to show bias towards a particular gender, race or sexual orientation simply based on the bias that was in the training data. I enumerate a couple of these examples. A model predicting the potential of a person to be a future CEO ended up selecting candidates who were white males, perhaps because till recently, CEOs of most top companies have been white males? An image recognition algorithm confused African faces with those of apes. Another model trained to identify good recruits from an employee pool showed bias against minorities and women. Needless to say, these embarrassments of models (some from reputed companies) came to a quick end.

Every nation including our democracy has had chapters in its existence we are not proud of. The data in agency databases contain the biases we exhibited as our society evolved. These biases against women, minorities and sexual orientation are inherent in the historical data we have. While developing a sound AI model, we must get over this data bias to avoid a biased model.

The top AI players like Google are working on tools to identify and rectify biased data. One way agencies can get started is with anonymizing training data by removing gender, race and other fields before training a model.

2. Malicious data

The last two years have shown how other countries have tried to influence US public opinion and policymaking by maliciously corrupting our media and national dialog. Since data is the building block of AI, corrupting the data foundation can corrupt the AI model as well. A couple of years ago, Microsoft released a chatbot that interacted intelligently with users and sought to adjust its behavior based on the interactions. A cohort of ill-intentioned users hit the bot with curses and filthy, racist language. Soon the bot was spewing venom against racial groups and different sexual orientations. Needless to say, Microsoft took the bot down with apologies but this was an excellent example of how malicious data can corrupt a model.

Summary

AI is going through a hype spike. Agency leaders and decision-makers are inundated with articles in the media as well as salespeople hyping their wares. AI is just a collection of advanced statistical methods and cybernetics that operate on the data that you have. We cut through the hype and dealt with some of the basics as well as terms that AI practitioners use. Since it involves software, AI is subject to all the rules and regulations agencies have to follow with software and systems. Agencies have plentiful data but challenges abound such as data silos and poor quality of data. This article points out themes for success and areas of opportunity to get started. Good luck with your advanced statistical journey and may the force be with you.

About the author

Sukumar Iyer is the founder and CEO of Brillient Corporation, a fast-growing federal contractor that specializes in big data analytics, AI, intelligent automation and business process management. Mr. Iyer has more than 26 years of experience in management consulting, information technology and software with a track record of successful projects, exceeding client expectations and delivering innovative solutions to challenging business problems.

On LinkedIn, on Twitter @SukumarIyer