There Is No AI Without IA

Seth Earley, *Earley Information Science*

Artificial intelligence (AI) is increasingly hyped by vendors of all shapes and sizes—from well-funded startups to the well-known software brands. Financial organizations are building AI-driven investment advisors.[1] Chat bots provide everything from customer service[2] to sales assistance.[3] Although AI is receiving a lot of visibility, the fact that these technologies all require some element of knowledge engineering, information architecture, and high-quality data sources is not well known. Many vendors sidestep this question or claim that their algorithms operate on unstructured information sources, "understand" those sources, interpret the user query, and present the result without predefined architectures or human intervention. That very well might be true in certain circumstances, but most applications require a significant amount of hard work on the part of humans before neural nets, machine learning, and natural language processors can work their magic.

DigitalGenius, a product that received press coverage at a confer-

ence in 2015,[4] uses deep learning and neural nets. DigitalGenius first classifies the incoming questions into one of several categories for further processing: product information, account information, action request, comparison question, recommendation question, and so on. These classifications qualify as a foundational element of information architecture. The starting point is contextualizing the query, which is then passed to other modules, including product information systems and other databases and APIs. Each of these systems and sources in turn needs to be well architected to return the correct information. If the information is not structured and curated in some way, there is nothing for the system to return. The ability of DigitalGenius to use AI to engage customers is predicated on having high-quality, structured data.

Digital Engagement: The Right Information in Context

Organizations are in a never-ending cycle of improving their digital means of engaging with

customers. These initiatives include improving personalization of the user experience by presenting more relevant content; tuning search results to return exactly what the user is interested in; and improving the effectiveness of offers and promotions. Organizations might also strive to increase response rates to email communications; provide better customer self-service; increase participation in user communities and other social media venues; and generally enhance the product experience through various other online mechanisms. In each of these cases, the means of engagement is providing a relevant piece of data or content (in the form of promotions, offers, sales, next-best action, products for cross sell and upsell, answers to questions, and so on) at exactly the right time and in the context that is most meaningful and valuable to the user.

This is done by interpreting the various signals that users provide through their current and past interactions with the organization. These signals include prior purchases, real-time click-stream

data, support center interactions, consumed content, preferences, buying characteristics, demographics, firmographics, social media information, and any other "electronic body language" that is captured by marketing automation and integration technologies. A search query could return a different result for a technical user than a nontechnical user, for example. At its core, search is a recommendation engine. The signal is the search phrase, and the recommendation is the result set. The more that is known about the user, the more the recommendation can be tailored. If the recommendation is related to a product, then clean, well-structured product data is a precondition.

Personalization, User Signals, and Recommendations

Getting these recommendations right and truly personalizing the user experience requires that product data be correctly structured and organized, that content processes be integrated into product onboarding, and that associations can be made among products, content, and user intent signals. The relationship of products to content is based on knowledge of the user's task and what content would help him or her accomplish it. The task might be a review, a how-to, product specifications, reference materials, instructions, diagrams and images, or other content that helps the user decide on the purchase.

AI encompasses a class of application that allows for easier interaction with computers and also allows computers to take on more of the types of problems that were typically in the realm of human cognition. Every AI program interfaces with information, and the better that information is structured, the more effective the

program is. A *corpus* of information contains the answers that the program is attempting to process and interpret. Structuring that information for retrieval is referred to as *knowledge engineering* and the structures are called *knowledge representation*.

Ontologies as Knowledge Representations

Knowledge representation consists of taxonomies, controlled vocabularies, thesaurus structures, and all of the relationships between terms and concepts. These elements collectively make up the *ontology*. An ontology represents a domain of knowledge and the information architecture structures and mechanisms for accessing and retrieving answers in specific contexts. Ontologies

can also capture "common sense" knowledge of objects, processes, materials, actions, events, and myriad more classes of real-world logical relationships. In this way, an ontology forms a foundation for computer reasoning even if the answer to a question is not explicitly contained in the corpus. The answer can be inferred from the facts, terms, and relationships in the ontology. In a practical sense, this can make the system more user friendly and forgiving when the user makes requests using phrase variations, and more capable when it encounters use cases that were not completely defined when it was developed. In effect, the system can "reason" and make logical deductions.

Correctly interpreting user signals enables the system to pres-

ent the right content for the user's context, and requires not only that our customer data is clean, properly structured, and integrated across multiple systems and processes but also that the system understand the relationship between the user, his or her specific task, the product, and the content needed—all assembled dynamically in real time. Building these structures and relationships and harmonizing the architecture across the various back-end platforms and front-end systems results in an enterprise ontology that enables a personalized, omnichannel experience. Some might call this an enterprise information architecture; however, there is more to it than the data structures. Recall that the definition of an ontology includes real-world

> Every AI program interfaces with information, and the better that information is structured, the more effective the program is.

logic and relationships. The ontology can contain knowledge about processes, customer needs, and content relationships.[5]

Mining Content for Product Relationships

Consumer and industrial products need to be associated with content and user context, but content can also be mined to suggest products for a user context. For an industrial application, a user might need parts and tools to complete maintenance on a hydraulic system. Using adaptive pattern-recognition software to mine reference manuals about hydraulic systems and repair, the system can extract a list of needed tools and related parts. A search on hydraulic repair will present a dynamically generated product

page based on the product relationships and correlated with the company's offerings. To some information professionals, this might sound complex and cumbersome to implement; however, there are emerging approaches that bring these aspirations closer to reality.

If It Works, It's Not AI

The concept of what constitutes AI has evolved as technology has evolved. A colleague of mine has said, "It's artificial intelligence until you know how it works." An interesting perspective indeed. I found support for this in material from an MIT AI course (see http://bit.ly/1WTCUXK):

There's another part of AI [...] that's fundamentally about applications. Some of these applications you might not want to call "intelligent" [...] For instance, compilers used to be considered AI, because [...] statements [were in a] high-level language; and how could a computer possibly understand? [The] work to make a computer understand [...] was taken to be AI. Now [...] we understand compilers, and there's a theory of how to build compilers [...] well, it's not AI anymore. [...] When they finally get something working, it gets co-opted by some other part of the field. So, by definition, no AI ever works; if it works, it's not AI.

Seemingly intractable problems have been solved by advances in processing power and capabilities. Not long ago, autonomous vehicles were considered technologically infeasible due to the volume of data that needed to be processed in real time. Speech recognition was unreliable and required extensive speaker-dependent training sessions. Mobile phones were once "auto-mobile" phones, requiring a car trunk full of equipment (my

first car phone in the '80s, in addition to costing several thousand dollars, left little room for anything else in the spacious trunk). Most AI is quietly taken for granted today. The word processor I am using was once considered an advanced AI application!

Simplicity Is Hidden Complexity

Under the covers, AI is complex; however, this complexity is hidden from the user and is, in fact, the enabler of an easy, intuitive experience. It is not magic, and it requires foundational structures that can be reused across many different processes, departments, and applications. These structures are generally first developed in silos and standalone tools; however, the true power will be realized when considered in a holistic framework of machine-intelligence-enabled infrastructure. AI will change the business landscape but will require investments in product and content architecture, customer data, and analytics, and the harmonization of tools in the customer engagement ecosystem. The organizations that adopt these approaches will gain significant advantages over the competition.

Clean Data Is the Price of Admission

AI approaches are proposed as the answer to enterprise challenges around improving customer engagement by dealing with information overload. Before those approaches can be leveraged, however, organizations need to possess the data needed as inputs to machine learning algorithms that can in turn process these diverse signals from unstructured and structured sources. Clean, well-structured, managed data is assumed.

In many cases, the data being processed or corpus being analyzed by AI systems is typically less structured than better-organized sources such as financial and transactional data. Learning algorithms can be used to both extract meaning from ambiguous queries and attempt to make sense of unstructured data inputs. Humans might phrase questions using different terminology, or they might ask overly broad questions. They are not always clear about their objectives—they don't necessarily know what they are looking for. This is why human sales people typically engage prospects in conversations about their overall needs, rather than asking them what they are looking for. (At least, the good ones do.)

Inserting AI into the process is more effective when users know what they want and can clearly articulate it, and where there is a relatively straightforward answer. The algorithm does what it does best in terms of handling variations in how those questions are asked, interpreting the meaning of the question, and processing other unstructured signals that help further contextualize the user's intent. There are many flavors of AI and many classes of algorithm that comprise AI systems. However, even when AI systems are used to find structure from completely unstructured information, they still require structure at the data layer.

Given that the data being searched by the AI system is unstructured, why do we need information architecture? Unstructured information is typically in the form of text from pages, documents, comments, surveys, social media, or some other source. Though it is unstructured, there are still parameters associated with

Table 1. Example applications for AI technology.

Application/problem	Typical approach	Why it will not work	How AI helps
Omnichannel selling	Integrations across systems that are inflexible	Content is typically out of synch, and systems have brittle integrations.	Harvesting signals of user intent across channels and devices; intelligent integration across diverse classes of customer-engagement technologies.
Personalization and recommendation engines	Classification of users according to personas and presentation of content based on use cases	Assumptions are rudimentary and simplistic, limit user choice, and rarely anticipate needs.	Incorporating multiple data sources, ontologies, and reasoning algorithms.
Dynamic content assembly	Limited assembly based on curated content	The number of permutations cannot be tracked, managed, or understood.	Assembling content that meets a user's preferences while incorporating real-time data, such as news and market performance supported by knowledge embedded in ontologies.
Field service equipment maintenance	View of maintenance schedules and analysis of field reports.	The volume of information from sensors and hidden signals from degrading performance is not perceivable by humans.	Processing performance data to predict or prevent maintenance events.
Customer self-service	Knowledge bases and FAQs	There are insufficient considerations of user context, language variations, fast-changing content, and a high volume of possible use cases.	Enabling intelligent agent knowledge base access using multiple variables to provide specific answers with metrics-driven feedback loops.
Call center support	Reps trained through on-the-job experience, knowledge bases	For technical and specialized knowledge, the level of experience needed is too costly.	Enabling intelligent agent access integrating answers and content from multiple knowledge and data sources using natural language interfaces.
Product and tool presentation	Hand-built landing pages	Maintenance of hand-curated and assembled landing pages becomes costly and unmanageable.	Harvesting product relationships through knowledge acquisition and mining of reference materials based on maintenance tasks and repair use cases.

the source and context. Social media information requires various parameters to describe users, their posts, relationships, time and location of posts, links, hashtags, and so on. The information architecture question in this case characterizes the structures of the input data, so that the system can be programmed to find patterns of interest. Even in the case of unsupervised machine learning (a class of application that derives signals from data that are not predefined by a human), the programmer still needs to describe the data in the first place with attributes and values. There might not be predefined categories for outliers and patterns that are

identified, but the input needs to have structure.

A common fallacy in considering big data sources that form inputs for machine learning is that because the data is "schema-less" (does not have a predefined structure), no structure is required. Data still requires attribute definitions, normalization, and cleansing to apply machine learning and pattern-identification algorithms.[6] As enterprises embark on the path to machine learning and AI, they should, first and foremost, be developing an enterprise ontology that represents all of the knowledge that any AI system they deploy would process, analyze, leverage, or require.

Some vendors might debate the value of this approach, insisting that their algorithms can handle whatever you throw at them—however, I would argue that this is only the case when the ontologies are self-contained in the tool. Even so, there will always be gaps between what a tool developed for broad adoption can contain and the specialized needs of an enterprise. Even if the tool is developed for a specific industry, the differences in processes found from organization to organization require specialized vocabularies and contextual knowledge relationships. This is a significant undertaking; however, not doing so misses an essential step in the process.

Table 2. Examples of AI tools with representative applications, limitations, considerations, and data sources.

Class of tool	Best suited for	Limitations
Inference engine	Deriving product and data relationships from unstructured content	Significant effort required in certain cases for development of adaptive pattern-recognition algorithms
Intelligent agent	Highly selective search and information retrieval for definable processes	Breadth of possible use cases, variability of queries and questions, degree of manual term-relationship mapping
Auto-classifier	Large volumes of higher-quality content where training sets or clear rules can be applied	Rules bases can become very complex, high volumes of training content required, variability of content
Entity extractors	Predictable formats for data (Social Security numbers, addresses, names, phone numbers, account numbers)	Ambiguity of entity values (for "Washington," can be Washington State, George Washington, Washington, DC, and so on), variations in format, quality of content
Unsupervised machine learning	Pattern detection, recognition, and prediction; anomaly detection, outliers, hidden attributes, and relationships; discovering new patterns or segmenting audiences, content, or data into clusters and groupings	Depends on creation of foundational hypothesis to define type of outlier or pattern (though not necessarily details of pattern); selection of class of algorithm requires technical sophistication and iterative testing across many classes of unsupervised learning from data signals to content, images, or transactions
Supervised machine learning	Pattern detection based on training data and used for well-understood patterns; discovering data, content, and relationships based on examples; finding similar documents, purchase patterns, audiences with like characteristics; predicting outcomes	Requires training sets and example data that teach the algorithm what type of information is being sought; needs large enough samples to test various hypotheses and an understanding of the specific outcome sought; risk of finding patterns where none exist (over fitting)
Hybrid unsupervised and supervised learning	Finding hidden patterns (unsupervised); using those patterns to train an algorithm to locate more instances of the data or content	Similar in concept to creating a survey with open-ended questions, then taking the answers to the survey and using those results to form closed-ended surveys

Much of what is described as AI is an extension of well-known approaches to addressing information management problems, all of which require clean foundational data and information structures as a starting point. The difference between standard information management and practical AI lies in understanding the limits of these technologies and where they can best be applied to address enterprise challenges.

The remainder of this article describes how your organization can identify use cases that will benefit from AI, identifies data sources that offer reliable and meaningful insights to train and guide AI, and defines governance, curation, and scalable processes that will allow for continuous improvement of AI and cognitive computing systems.

Identifying Use Cases

Differentiating AI use cases from standard information management use cases requires considering the sources of data that comprise the "signals" being processed, the type of task the user faces, and the systems that will be part of the solution. The difference in approaches to these problems lies in how the sources of data will be curated and ingested, how organizing principles need to be derived and applied, the sophistication of the functionality desired, and the limitations of the current solution in place. An

AI approach will require a greater level of investment, sponsorship at the executive level, program-level governance, and an enterprise span of influence. It will also require a longer-term commitment than a typical information management project. Although there are opportunities to deploy limited scope AI, fully leveraging it as a transformative class of technology should be part of an overall digital transformation strategy in some cases on the scale of enterprise resource planning (ERP) programs, with commensurate support, funding, and commitment. (Some ERP programs can cost US$50 to $100 million or more). Although no organization will approach an unproven set of technologies with

Decision factors	Example data
Volume of content, quality of corpus, structure of data within content (use of tables, headers, other identifiers, and delimiters)	Structured documents such as catalogs, engineering specifications, knowledge bases, ERP data
Narrow scenarios, defined repeatable tasks, sophisticated audience, standard terminology for domain	Policy and procedure manuals, chat logs, customer service databases, support tickets, knowledge harvested from human subject matter experts
High-value content that has a predictable structure, curated with editorial standards	Semi-structured content such as regulations, journal articles, news articles, policies and procedures, support documentation, engineering documents
Security, compliance, and reference-based applications (looking for personally identifiable information, sensitive account information, intellectual property protection, fact identification)	Same semi-structured content sources mentioned above, along with controlled vocabularies and thesaurus structures to identify entity values
Determination of type of patterns that are being detected; not typically selected in isolation because learning algorithms are embedded in many aspects of AI application; requires guidance of data science specialist knowledgeable in specific type of algorithm	Unsupervised learning can be applied to virtually any data source as the tools look for patterns; can be applied to unstructured and semi-structured content as well as transactional data, images, audio and video, scientific datasets, sensor data, social media sources, and so on
Similar challenges to unsupervised learning: typically embedded in AI applications rather than being applied as standalone tool and requires specialized knowledge; can be used to find types of images, content or data, purchase patterns, audiences, users, combinations of products, or other patterns depending on data sources	Same as unsupervised; the difference is having training sets that contain patterns sought
Most AI programs contain hybrid implementations of supervised and unsupervised and therefore the complexity of each approach can be compounded; requires research and experimentation to find correct combinations and fine-tune algorithms	Same data used for unsupervised learning can be used as input for supervised learning after representative patterns have been identified

that level of commitment, funding needs to be allocated to extend proven approaches with emerging AI technologies.

The roadmap for AI transformation includes continuous evaluation of payback and ROI and focuses on short-term wins while pursuing longer-term goals. Most organizations are attempting to solve the problems described in Table 1 with limited approaches, departmental-level solutions, standalone tools, and insufficient funding. These problem classes are facing most enterprises, and though progress can be made with limited resources and siloed approaches, this would be an extension of business as usual. Truly transformative applications will require an enterprise view of the organization's knowledge landscape and implementation of new governance, metrics, and data quality programs—governance to make decisions, metrics to monitor the effectiveness of those decisions, and data quality to fuel the AI engine.

Table 1 presents example applications for AI technology.

Identifying Data Sources

Training data can come from typical knowledge bases, the more highly curated the better. Call center recordings and chat logs can be mined for content and data relationships as well as answers to questions. Streaming sensor data can be correlated with historical maintenance records, and search logs can be mined for use cases and user problems. Customer account data and purchase history can be processed to look for similarities in buyers and predict responses to offers; email response metrics can be processed with text content of offers to surface buyer segments. Product catalogs and data sheets are sources of attributes and attribute values. Public references can be used for procedures, tool lists, and product associations. YouTube video content audio tracks can be converted to text and mined for product associations. User website behaviors can be correlated with offers and dynamic content. Sentiment analysis, user-generated content, social graph data, and other external data sources can all be mined and recombined to yield knowledge and

user-intent signals. The correct data sources depend on the application, use cases, and objectives.

Table 2 describes examples of AI tools with representative applications, limitations, considerations, and data sources. Though not meant to be an exhaustive list, and recognizing that one class of tool is frequently leveraged in other tools and applications (an intelligent agent can use inference engines, which in turn can leverage learning algorithms, for example), the table articulates considerations for exploring one approach versus another.

Defining Governance, Curation, and Scalable Processes

AI and cognitive computing are managed in the same way as many other information and technology governance programs. They require executive sponsorship, charters, roles and responsibilities, decision-making protocols, escalation processes, defined agendas, and linkage to specific business objectives and processes. These initiatives are a subset of digital transformation and are linked to customer life cycles and internal value chains. Because the objective is always to affect a process outcome, all AI and cognitive computing programs are closely aligned with ongoing metrics at multiple levels of detail—from content and data quality to process effectiveness and satisfaction of business imperatives—and ultimately are linked to the organizational competitive and market strategy. Milestones and stages are defined to release funding for program phases, each with defined success criteria and measurable outcomes.

AI will no doubt continue to impact every aspect of our personal and professional lives. Much of this impact will occur in subtle ways—such as improved usability of applications and findability of information. These will not necessarily appear on the surface to be AI. Over time, AI-driven intelligent virtual assistants will become more fluent and capable, and will become the preferred mechanism for interacting with technology. Humans create knowledge, while machines process, store, and act on that knowledge. AI is applied human knowledge. Organizations need to develop the foundations for advancing AI by capturing and curating that knowledge and by building the foundational data structures that form the scaffolding for that knowledge. Without those components, the algorithms have nothing to run on. **IT**

References

1. J. Vögeli, "UBS Turns to Artificial Intelligence to Advise Clients," Bloomberg, 7 Dec. 2014; www.bloomberg.com/news/articles/2014-12-07/ubs-turns-to-artificial-intelligence-to-advise-wealthy-clients.
2. C. Green, "Is Artificial Intelligence the Future of Customer Service?" MyCustomer, 3 Dec. 2015; www.mycustomer.com/service/channels/is-artificial-intelligence-the-future-of-customer-service.
3. E. Dwoskin, "Can Artificial Intelligence Sell Shoes?" blog, *Wall Street J.*, 17 Nov. 2015; http://blogs.wsj.com/digits/2015/11/17/can-artificial-intelligence-sell-shoes/.
4. R. Miller, "DigitalGenius Brings Artificial Intelligence to Customer Service via SMS," Tech Crunch, 5 May 2015; http://techcrunch.com/2015/05/05/digitalgenius-brings-artificial-intelligence-to-customer-service-via-sms/.
5. S. Earley, "Lessons from Alexa: Artificial Intelligence and Machine Learning Use Cases," blog, Earley Information Science, 24 Mar. 2016; www.earley.com/blog/lessons-alexa-artificial-intelligence-and-machine-learning-use-cases.
6. J. Brownlee, "How to Prepare Data for Machine Learning," Machine Learning Mastery, 25 Dec. 2013; http://machinelearningmastery.com/how-to-prepare-data-for-machine-learning/.

Seth Earley is CEO of Earley Information Science (www.earley.com). He's an expert in knowledge processes, enterprise data architecture, and customer experience management strategies. His interests include customer experience analytics, knowledge management, structured and unstructured data systems and strategy, and machine learning. Contact him at seth@earley.com.

cn Selected CS articles and columns are available for free at http://ComputingNow.computer.org.

ITPro