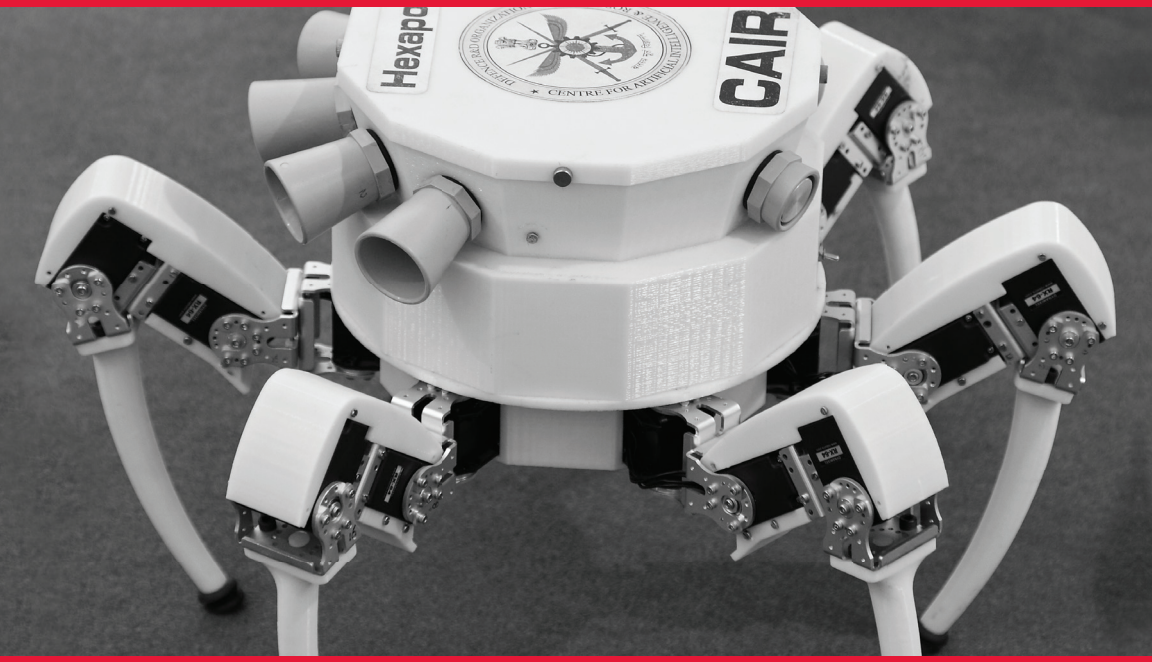
O'REILLY®

Practical Artificial Intelligence in the Cloud

Exploring AI-as-a-Service for Business and Research

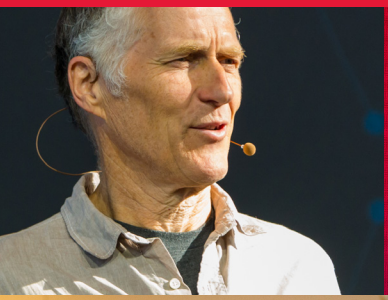


Mike Barlow

AI is moving fast. Don't fall behind.

Early adopters of applied AI have a unique opportunity to invent new business models, reshape industries, and build the impossible.

Put AI to work—right now.





Artificial Intelligence CONFERENCE

theaiconf.com

PRESENTED BY

O'REILLY® intel Nervana™

Practical Artificial Intelligence in the Cloud

Exploring AI-as-a-Service for Business and Research

Mike Barlow

Beijing • Boston • Farnham • Sebastopol • Tokyo O'REILLY®

Practical Artificial Intelligence in the Cloud
by Mike Barlow

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://safaribooksonline.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Acquisitions Editor: Shannon Cutt
Editor: Shannon Cutt
Production Editor: Melanie Yarbrough
Interior Designer: David Futato
Cover Designer: Randy Comer
Illustrator: Rebecca Demarest

August 2016: First Edition

Revision History for the First Edition

2016-08-31: First Release
2016-10-19: Second Release

978-1-491-96739-3

[LSI]

Table of Contents

Practical Artificial Intelligence in the Cloud

When the automobile was introduced, there were rumors that drivers and passengers would suffocate when the speed of their vehicles exceeded 20 miles per hour. Many people believed that cars would never become popular because the noise of passing engines frightened horses and other farm animals.

Nobody foresaw rush-hour traffic jams, bumper stickers, or accidents caused by people trying to text and drive at the same time.

It's hard to imagine AI (artificial intelligence) spooking farm animals. But that hasn't stopped several generations of science-fiction writers from inventing scary stories about the rise of sentient computers and killer robots.

We can't see the future, and it's impossible to predict with any reasonable degree of accuracy how AI will change our lives. But we can make some educated guesses. For instance, it seems clear that AI as a global phenomenon is growing rapidly, and that a large piece of that growth is enabled by the cloud.

As a society, we're no longer debating whether AI is feasible or practical. Instead, we're asking where, when, and how AI can be used to solve problems, achieve higher levels of efficiency, apply knowledge more effectively, and improve the human condition.

What is increasingly apparent is that the sizes of the applications and datasets required for genuine AI processes are too large for devices such as smart phones or laptops. The idea of AI running

independently on local machines evokes images of early factories that generated their own electrical power.

To be fair, it's likely that small devices will eventually have enough processing power and data storage capacity to run AI programs "off the grid," but that day is still far in the future. For the meantime, we'll need the cloud to take advantage of AI's potential as a tool for progress.

Old Categories Vanishing

Back in the days when AI was seen as something wildly "futuristic," science writers tended to lump it into three broad categories:

1. Narrow (Weak AI)
2. Human-level (Strong AI)
3. Smarter-than-human (Superintelligence)

Weak AI was often portrayed as puny and useless. Strong AI was "just over the horizon" or "several years down the road." Superintelligence, a term credited to Oxford philosopher Nick Bostrom, refers to "an intellect that is much smarter than the best human brains in practically every field, including scientific creativity, general wisdom, and social skills." As far as anyone knows, superintelligence doesn't exist—but that hasn't stopped respected intellectual celebrities like Elon Musk and Stephen Hawking from issuing warnings about its apocalyptic potential.

For better or worse, the emergence of many commercially produced AI products and services has rendered those categories largely irrelevant. In this report, I'll be writing about "Practical AI," a term I've coined to describe the kinds of AI we're already using or likely to be using within the next six months.

For the purpose of this report, Practical AI includes related techniques such as machine learning, neural networks, deep learning, text analytics, classification, visual recognition, and NLP (natural language processing).

Here are the top takeaways from my interviews with experts from organizations offering AI products and services:

- AI is too big for any single device or system
- AI is a distributed phenomenon

- AI will deliver value to users through devices, but the heavy lifting will be performed in the cloud
- AI is a two-way street, with information passed back and forth between local devices and remote systems
- AI apps and interfaces will be designed and engineered increasingly for nontechnical users
- Companies will incorporate AI capabilities into new products and services routinely
- A new generation of AI-enriched products and services will be connected and supported through the cloud
- AI in the cloud will become a standard combination, like peanut butter and jelly

"It's inevitable that AI will move into the cloud," says Nova Spivack, CEO and cofounder of Bottlenose, a business intelligence software company. Spivack is the author of "Why Cognition-as-a-Service is the next operating system battlefield," an article in which he makes the case for on-demand AI.

"If you're talking about systems that have to analyze hundreds of billions of data points continuously and run machine learning models on them, or do difficult things like natural language processing and unstructured data mining—those processes require a lot of storage, a lot of data, a lot of computation," he says. "So it makes sense to centralize them in the cloud. But there will also be situations requiring hybrid approaches that leverage local processors and devices."

Cloud-based AI products and services are easier to update than onpremise versions, says Naveen Rao, CEO and cofounder of Nervana Systems, a company that offers AI-as-a-service through Nervana Cloud. The company recently agreed to be acquired by Intel. "We're constantly developing, adding new features, and updating our products. If everything is taking place within your existing infrastructure, it becomes very difficult to add those new features and updates," he says.

While the idea of ceding control of AI infrastructure to vendors might not appeal to some customers, the alternative can be equally unappetizing. "There's been a lot of talk about trying to make AI work on existing infrastructure," says Rao. "But the sad reality is that you're always going to end up with something that's far less than state-of-the-art. And I don't mean it will be 30 or 40 percent slower. It's more likely to be a thousand times slower."

With cloud-based AI, you can "mix and match" the latest technologies and the most advanced techniques. "We're at the point where we have much better building blocks. It's like going from older DUPLO blocks to newer, fancier LEGO blocks. Today we have a whole new set of pieces you can assemble in new ways to build really cool new things," says Rao.

The cloud will also accelerate the democratization of AI and other advanced analytics, says Mark Hammond, CEO and founder of Bonsai, a company that "makes it easy for every developer to program artificial intelligence" applications and systems.

"There are 18 million developers in the world, but only one in a thousand have expertise in artificial intelligence," he says. "To a lot of developers, AI is inscrutable and inaccessible. We're trying to ease the burden."

If Bonsai's mission succeeds, it will do for AI "what Visual Basic did for desktop applications, what PHP did for the first generation of web applications and what Ruby on Rails did for the next generation of web applications," says Hammond.

"We're doing for AI what databases did for data," he says. "We're trying to abstract away the lower levels and common concerns. Nobody wants their company's core competency to be managing data in a database. We feel the same way about AI."

In many ways, Hammond represents a wave of entrepreneurs who are counting on the cloud to help them make AI less exotic and more accessible. That's bad news for science-fiction writers and AI doomsayers, but good news for the rest of us.

Powering Economic Transformation

Thanks to a perfect storm of recent advances in the tech industry, AI has risen from the ashes and regained its aura of cool. Two years ago, AI was a cliché, a sad remnant of 1950s-style futurism. Today it's sexy again. Most large software vendors now offer suites of AI products and services available through the cloud. They're not merely jumping on the bandwagon—they're convinced that AI will become a major force in the economy.

"There isn't a single industry that won't be transformed," says Rob High, vice president and chief technology officer for IBM Watson. "We can literally build cognition into everything digital."

For example, Watson technology has already been applied to medical research, oil exploration, educational toys, personal fitness, hospitality, and complex financial systems.

IBM recently announced a collaborative deal with Twilio, a cloud communications platform used by more than one million developers. As part of the collaboration, IBM introduced two new offerings, IBM Watson Message Sentiment and IBM Watson Message Insights. Both are available as Add-Ons in Twilio Marketplace.

"We're focused on building out the most extensive cognitive capabilities in an open platform, including the areas of speech, language, and vision," says High.

Google has also thrown its hat into the ring. "I'm excited to see the rising tide of innovation that will come out of machine learning," says Fausto Ibarra, director of product management, data analytics, and cloud machine learning for the Google Cloud Platform.

"One of my favorite examples is a developer who used our Cloud Vision API and Cloud Speech API to create an app that helps blind and visually impaired users identify objects," says Ibarra. "City governments in Europe and Asia are using data from road sensors with machine learning to optimize traffic flows and dramatically increase the efficiency of public transportation."

Machine learning, he says, is becoming an essential element in applications across many industries. In an effort to make machine learning more accessible, Google open sourced TensorFlow, a framework that gives developers access to core technologies that Google uses to bring machine intelligence into its own services.

"Since we introduced TensorFlow, it has become the most popular machine learning project on GitHub," says Ibarra.

Google is steadily pushing forward with its cloud-based AI ecosystem. For example, developers can use TensorFlow Serving with Kubernetes to scale and serve machine learning models. In July 2016, Google launched a beta version of Cloud Natural Language API, a machine learning product that can be used to reveal the structure and meaning of text in a variety of languages.

Additionally, Google has "partnered with a number of organizations, including data Artisans, Cloudera, Talend, and others to submit the Dataflow model, SDKs, and runners for popular OSS distributed systems to the Apache Incubator. This new incubating project, known as Apache Beam, allows you to define portable, powerful, and simple data processing pipelines that can execute in either streaming or batch mode," according to a recent post in the Google Cloud Platform Blog.

A Full Menu of APIs for AI

Most of the world's large software vendors have committed to playing in the AI space. Hewlett Packard Enterprise (HPE), for example, has launched HPE Haven OnDemand, "a platform for building cognitive computing solutions using text analysis, speech recognition, image analysis, indexing and search APIs".

HPE Haven OnDemand offers free APIs for audio-video analysis, geo analysis, graph analysis, image analysis, format conversion, and unstructured text indexing. As the needs of AI developers evolve, the menu of APIs evolves, too. Within audio-video analysis, for example, are APIs for detecting changes in scenes and recognizing license plates.

"Haven OnDemand is all about applied machine learning," says Fernando Lucini, chief technology officer for big data at HPE. "It's a self-service platform in the cloud."

From Lucini's perspective, companies like HPE have already made significant strides in transforming AI from a mysterious black box into a user-friendly set of tools.

"In the past, you would have done massive amounts of planning. You would have worried about budgets and people. Now the only question is whether you have the hunger to get started," he says.

For instance, let's say you want to analyze 100 gigabytes of email (roughly 2 million messages) in hopes of gleaning insights that might lead to developing new products or new ways of managing information. "Who has the appetite to sift through 2 million pieces of email? Nobody, of course! You would go to the pub, and that would be the end of it," says Lucini.

With AI in the cloud, however, you would be able to access both the applications and the computing power necessary to sift through huge numbers of emails without breaking a sweat. "Now there is no barrier," says Lucini. "And when you're done, you just fold it up. To me, that's fundamentally exciting."

Building in the Cloud

Lucini foresees AI in the cloud penetrating multiple sectors of the broader economy. "I think all industries are going to take advantage of this. If you're crunching through huge amounts of data, the cloud is the only way to go," he says. "Why buy 1,000 machines to do a job when you can rent the machines in the cloud for a couple of weeks?"

David Laxer, a data scientist and software developer, says it's hard to justify purchasing "custom hardware that will be obsolete in a year or two" when you can rent or lease the resources you need in the cloud.

"Let's say I'm working on a semantic hashing algorithm and my document collection is huge—say, the size of the U.S. Patent Office database. I can't do that on my Mac. I have to do that in the cloud," says Laxer. "I can upload my data to an EC2 (Amazon Elastic Compute Cloud) instance and start training my models with deep learning using Spark, do the testing, and actually deploy an application."

Additionally, developers can choose among several options for renting compute resources in the cloud. "With EC2, for example, you can get a reserved arrangement where particular servers are yours for a month or for as long you need them. Or you can go a cheaper route and bid on what Amazon calls 'Spot instances.' The downside of bidding is that if someone outbids you, then you lose the instance. It's like buying a reserved seat at a ballpark versus buying a seat in the bleachers," he explains.

From the perspective of freelance AI developers, the cloud offers the best deal. "You can't do this in your garage. The cost of buying servers would be prohibitive," he says.

However, it can take a while for some advances in technology to fully penetrate the cloud. Not every cloud provider offers GPU instances, says Laxer, and some of the available GPU instances are several years old.

In some cases, it might make more sense to buy a high-performance board and install it in a heavy-duty workstation—assuming you have air conditioning or water cooling. But you can only put four boards into a server unless you've got a rack mount, and you're probably not going to do that in your garage, he says.

"If you're Google or Facebook, you've got buildings full of servers with GPUs," says Laxer. "If you're a freelance developer and you're training deep belief networks or doing reinforcement learning, you won't be able to solve AlphaGo-class problems".

Tangible Economic Benefits

The real-world experiences of Laxer and other developers demonstrate the value and practicality of AI in the cloud. As a business model, AI in the cloud offers tangible economic benefits for developers and cloud vendors.

It also offers opportunities for developers to have fun and enjoy the pleasures of working hands-on with cool new technology. For example, IBM's Watson Developer Cloud (WDC) makes it easy to program poker-playing robots with cognitive capabilities.

"The robots should be able to look at their cards and recognize them. They should listen to what the players around the table are talking about and understand it. They should be able to participate in conversations," writes Dhaval Sonawane, an IBM Watson intern. "In the past it has been difficult to do these tasks using conventional methods but Watson will help us accomplish these normally difficult tasks."

Five capabilities are required for creating poker-playing robots, writes Sonawane, and WDC provides all of them:

1. Speech to Text
2. Text to Speech
3. Natural Language Classifier
4. Dialog
5. Visual Recognition

For developers, poker-playing robots are just the tip of the AI iceberg. Josh Patterson, director of field engineering at Skymind, an open-source, enterprise deep-learning provider, foresees the emergence of a whole new galaxy of sophisticated cloud-based AI tools,

techniques, and applications for professional developers and entrepreneurs.

"Image classification is an area that will progressively rely on pre-trained models (augmented by custom models). Andrej Karpathy has shown exceptional results with Neural Talk, his line of research that produces captions for images. Users can download this model today and leverage it in their own applications," says Patterson. "I see this pattern of commoditizing image classification continuing to the point where a cloud application could rely entirely on a pre-built image model and never do any custom training."

He predicts that applications in audio and video could follow a similar pattern. "DL4J will soon offer the ability to load models from other deep learning tools (e.g., Caffe, TensorFlow), making it easier to productionize enterprise deep learning anywhere on the JVM (Java Virtual Machine)," he says.

Patterson also sees growth in the area of pre-built generalized models. "We're already seeing Apple building tools into iOS that apply convolutional models for images," he says. "This allows developers to apply saved models in their own applications."

Consumer-Facing Impact of AI in the Cloud

Developers aren't the only target market for AI in the cloud products and services. Consumers will feel the impact most strongly when they deal with customer service operations.

"AI will follow in the paths of previous strategies to reduce labor costs at call centers," says Kanishk Priyadarshi, a Silicon Valley tech executive. "But instead of degrading service, AI will actually improve it."

In the 1990s and early 2000s, widespread efforts to slash costs through labor arbitrage decimated call center operations across the consumer retail industry, damaging brands and ruining reputations. "There's little argument that labor arbitrage degraded customer service. It was catastrophic for customer satisfaction scores," says Priyadarshi. "AI can reverse that trend and raise levels of customer satisfaction in many industries."

With AI in the mix, consumers will either deal directly with specialized bots or with service representatives who are supported by AI

services. "Overall, the customer experience will improve because we'll have direct access to the best knowledge and the best expertise. The chat bots we engage with will have broad and deep knowledge of the products and services we're using," he says. "Our conversations with chat bots and digital agents will be highly personalized. And they'll already know about our problems, so every conversation won't have to start from scratch."

Moreover, AI-supported customer service operations will learn continuously from each customer interaction. "The bots will know that you've called five times and that you're frustrated. They'll have access to in-depth knowledge about everyone else who's called with a similar issue, so they'll know which answer is most likely to resolve your problem. And all of this will seem to happen instantaneously," says Priyadarshi.

"For years, we've been talking about big data. This is where big data finally becomes useful to large numbers of people. The robot will know the answer to your question before you even ask," he says.

Next IT, a major player in the "conversational AI" space, has built intelligent chat bots and virtual assistants for customers like the U.S. Army, Amtrak, AT&T, and Aetna. Next IT's innovative approach to the market represents a step in the elevation of commercial AI from a back-office tool for cutting costs to an indispensable driver of deeper and more effective customer engagement strategies.

"Chat bots and virtual assistants go beyond simply customer support," explains Rick Collins, president of Next IT's enterprise business. "They're driving a lot of value as trusted brokers of conversations between our customers and their customers. That's what's really exciting, when technology enables revenue in addition to cost reduction and productivity."

Intelligent virtual assistants such as "Julie" of Amtrak, "Ann" of Aetna, and "SGT STAR" of the U.S. Army aren't merely advanced customer service tools—they're the new "faces" of AI in the cloud. Unlike Apple's Siri and Microsoft's Cortana, which offer general information on a wide range of subjects, the commercial AI bots offer deep, specialized knowledge for customers with specific questions and issues.

In the near future, most companies will rely increasingly on AI-supported agents and bots to manage customer relationships, lead-

ing to a radical transformation of commerce. Instead of simply aiming for maximum efficiency at the lowest possible cost, the AI-enabled customer contact center of tomorrow will be seen as a profit center that keeps customers interested, engaged, and loyal.

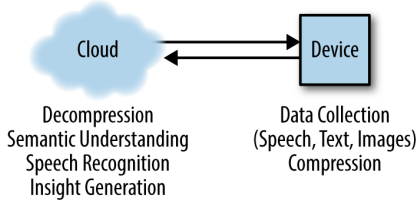
Evolving Partnerships between Clouds and Devices

Back in April, Movidius unveiled its Fathom Neural Compute Stick and NVIDIA announced its Pascal-based Tesla P100 GPU with HBM2. Both announcements demonstrate the evolutionary path of AI in the cloud. On the one hand, "traditional" vendors like Google, IBM, Microsoft, Amazon, and HPE are rolling out sophisticated cloud services for AI developers and users. On the other hand, newer vendors like NVIDIA and Movidius are radically enhancing the AI capabilities of devices and local networks.

PP Zhu, founder, president, and chief technology officer at Xiaoi Robot, a major producer of smart machine technology in China, sees AI in the cloud as a continuous interplay between local and remote technologies.

Data collection and compression are performed at the device level while advanced processes such as semantic analysis, natural language processing, and machine learning are performed in the cloud, says Zhu.

Some of the insights and knowledge created by the advanced processes is passed back to apps on the device, creating a dynamic feedback loop. "It's almost as if the cloud acts a teacher," says Zhu.



It's only a matter of time before the expertise and knowledge gained from developing smart bots and virtual customer assistants can be transferred into physical robots that interact directly with humans. Last year, Xiaoi launched a smart robot cloud operating system that

Zhu says can be embedded into appliances, vehicles, and a variety of devices, including physical robots.

What About the Shannon Limit?

Clearly, the dawn of the robots is approaching far more quickly than we had imagined. Many of Isaac Asimov's robot stories take place in the distant future. Thanks to AI in the cloud, we're likely to be working alongside robots within the next couple of years. Some of us already are partnering with robots, even when we're not directly aware of them.

The rapid acceleration of AI in the cloud raises important questions about the readiness of our data networks to handle enormous flows of critical information in real time. As I wrote in two previous reports, *Are Your Networks Ready for the IoT?* and *Evolving Infrastructures of the Industrial IoT*, we are fast approaching the Shannon Limit, the theoretical barrier limiting the amount of information a system can safely manage. Go beyond that limit and your data will degrade, according to the legendary information theorist Claude Shannon.

Until very recently, the Shannon Limit was considered an interesting footnote to the history of information theory. Today, the consequences of hitting the Shannon Limit would be very tangible: Imagine what might happen if the AI-supported driverless car taking you to work begins receiving bad data from the cloud because the Shannon Limit has been exceeded?

One obvious answer would be that the driverless car would need a certain amount of AI capability built into it. "The decision logic really has to be in the vehicle, because the cloud is not perfect," says Bryan Reimer, a research scientist in the MIT AgeLab and associate director of The New England University Transportation Center at MIT. "If you're in a tunnel or in the middle of a blank spot on the grid, it would be problematic."

How much AI needs to be embedded within the car or device and how much can be safely entrusted to the cloud? Who will decide the right proportion of cloud-based and embedded reasoning capabilities? Will those kinds of decisions be left to manufacturers, legislators, courts, or consumers?

Those kinds of questions will become increasingly material to further discussion about AI in the cloud and its broader impact on modern culture.

About the Author

Mike Barlow is an award-winning journalist, author, and communications strategy consultant. Since launching his own firm, Cumulus Partners, he has worked with various organizations in numerous industries.

Barlow is the author of *Learning to Love Data Science* (O'Reilly, 2015). He is the coauthor of *The Executive's Guide to Enterprise Social Media Strategy* (Wiley, 2011), and *Partnering with the CIO* (Wiley, 2007). He is also the writer of many articles, reports, and white papers on numerous topics including smart cities, ambient computing, predictive maintenance, advanced data analytics, and infrastructure.

Over the course of a long career, Barlow was a reporter and editor at several respected suburban daily newspapers, including *The Journal News* and the *Stamford Advocate*. His feature stories and columns appeared regularly in *The Los Angeles Times*, *Chicago Tribune*, *Miami Herald*, *Newsday*, and other major US dailies. He has also written extensively for O'Reilly Media.

A graduate of Hamilton College, he is a licensed private pilot, avid reader, and enthusiastic ice hockey fan.