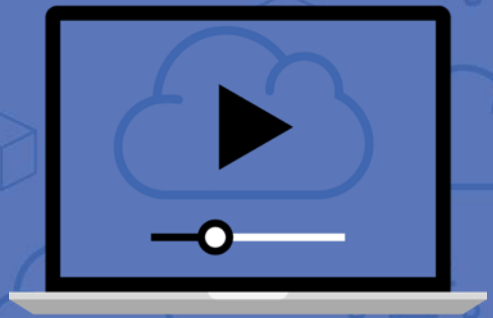
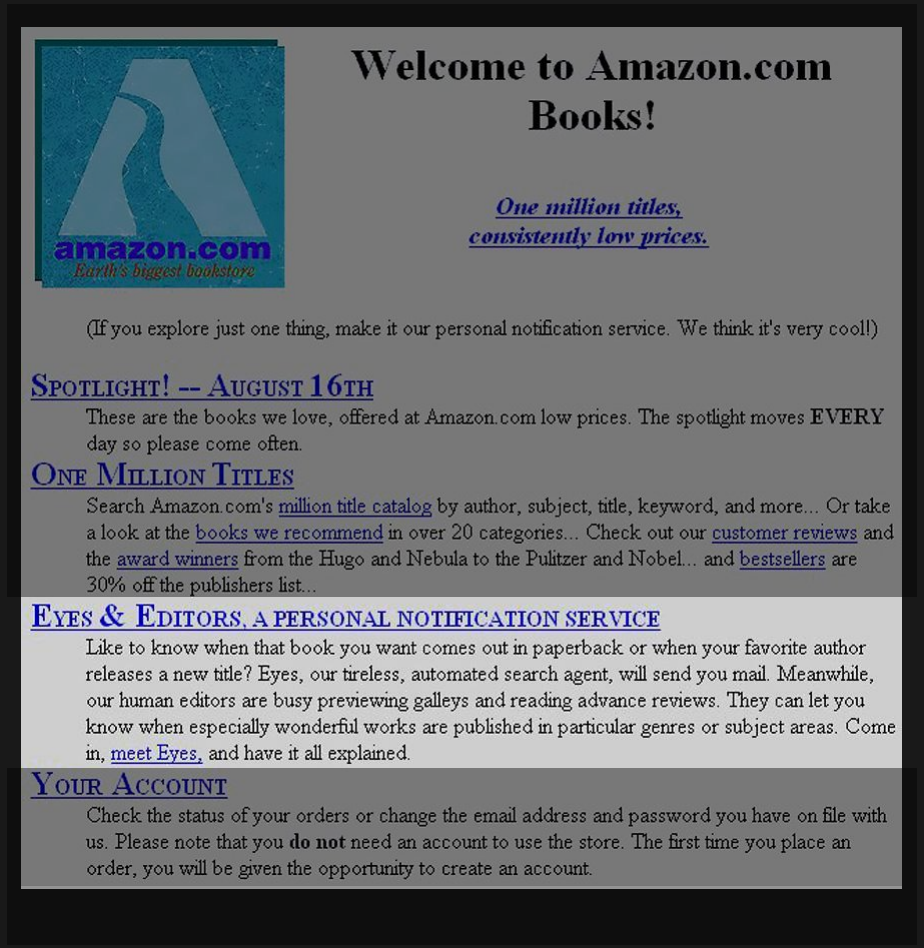
Machine Learning: From Notebook to Production with Amazon Sagemaker

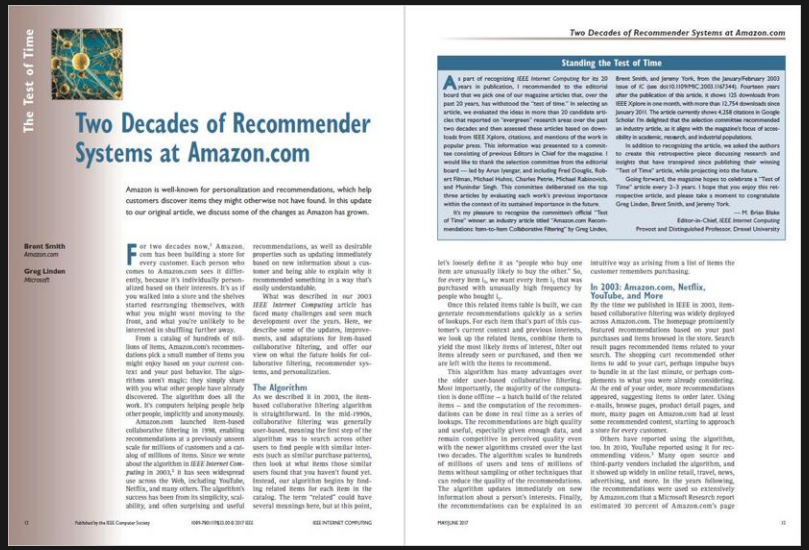


Julien Simon, AI Evangelist, EMEA
@julsimon



Amazon.com,1995

« *Two Decades of Recommender Systems at Amazon.com* » (2017)



The Test of Time

Two Decades of Recommender Systems at Amazon.com

Amazon is well-known for personalization and recommendations, which help customers discover items they might otherwise not have found. In this update to our original article, we discuss some of the changes as Amazon has grown.

Brent Smith
Amazon.com

Greg Linden
Microsoft

For two decades now,[1] Amazon.com has been building a store for every customer. Each person who comes to Amazon.com sees it differently, because it's individually personalized based on their interests. It's as if you walked into a store and the shelves started rearranging themselves, with what you might want moving to the front, and what you're unlikely to be interested in shuffling further away.

From a catalog of hundreds of millions of items, Amazon.com's recommendations pick a small number of items you might enjoy based on your current context and your past behavior. The algorithms aren't magic; they simply share with you what other people have already discovered. The algorithm does all the work. It's computers helping people help other people, implicitly and anonymously.

Amazon.com launched item-based collaborative filtering in 1998, enabling recommendations at a previously unseen scale for millions of customers and a catalog of millions of items. Since we wrote about the algorithm in *IEEE Internet Computing* in 2003,[2] it has seen widespread use across the Web, including YouTube, Netflix, and many others. The algorithm's success has been from its simplicity, scalability, and often surprising and useful recommendations, as well as desirable properties such as updating immediately based on new information about a customer and being able to explain why it recommended something in a way that's easily understandable.

What was described in our 2003 *IEEE Internet Computing* article has faced many challenges and seen much development over the years. Here, we describe some of the updates, improvements, and adaptations for item-based collaborative filtering, and offer our view on what the future holds for collaborative filtering, recommender systems, and personalization.

The Algorithm

As we described it in 2003, the item-based collaborative filtering algorithm is straightforward. In the mid-1990s, collaborative filtering was generally user-based, meaning the first step of the algorithm was to search across other users to find people with similar interests (such as similar purchase patterns), then look at what items those similar users found that you haven't found yet. Instead, our algorithm begins by finding related items for each item in the catalog. The term "related" could have several meanings here, but at this point,

Published by the IEEE Computer Society

1089-7801/17/$33.00 © 2017 IEEE

Standing the Test of Time

As part of recognizing *IEEE Internet Computing* for its 20 years in publication, I recommended to the editorial board that we pick one of our magazine articles that, over the past 20 years, has withstood the "test of time." In selecting an article, we evaluated the ideas in more than 20 candidate articles that reported on "evergreen" research areas over the past two decades and then assessed these articles based on downloads from IEEE Xplore, citations, and mentions of the work in popular press. This information was presented to a committee consisting of previous Editors in Chief for the magazine. I would like to thank the selection committee from the editorial board — led by Arun Iyengar, and including Fred Douglis, Robert Fillman, Michael Huhns, Charles Petrie, Michael Rabinovich, and Munindar Singh. This committee deliberated on the top three articles by evaluating each work's previous importance within the context of its sustained importance in the future.

It's my pleasure to recognize the committee's official "Test of Time" winner: an industry article titled "Amazon.com Recommendations: Item-to-Item Collaborative Filtering" by Greg Linden, Brent Smith, and Jeremy York, from the January/February 2003 issue of IC (see doi:10.1109/MIC.2003.1167344). Fourteen years after the publication of the article, it shows 135 downloads from IEEE Xplore in one month, with more than 12,754 downloads since January 2011. The article currently shows 4,358 citations in Google Scholar. I'm delighted that the selection committee recommended an industry article, as it aligns with the magazine's focus of accessibility in academic, research, and industrial populations.

In addition to recognizing the article, we asked the authors to create this retrospective piece discussing research and insights that have transpired since publishing their winning "Test of Time" article, while projecting into the future.

Going forward, the magazine hopes to celebrate a "Test of Time" article every 3–5 years. I hope that you enjoy this retrospective article, and please take a moment to congratulate Greg Linden, Brent Smith, and Jeremy York.

— H. Brian Blake
Editor-in-Chief, *IEEE Internet Computing*
Provost and Distinguished Professor, Drexel University

let's loosely define it as "people who buy one item are unusually likely to buy the other." So, for every item i_1, we want every item i_2 that was purchased with unusually high frequency by people who bought i_1.

Once this related items table is built, we can generate recommendations quickly as a series of lookups. For each item that's part of this customer's current context and previous interests, we look up the related items, combine them to yield the most likely items of interest, filter out items already seen or purchased, and then we are left with the items to recommend.

This algorithm has many advantages over the older user-based collaborative filtering. Most importantly, the majority of the computation is done offline — a batch build of the related items — and the computation of the recommendations can be done in real time as a series of lookups. The recommendations are high quality and useful, especially given enough data, and remain competitive in perceived quality even with the newer algorithms created over the last two decades. The algorithm scales to hundreds of millions of users and tens of millions of items without sampling or other techniques that can reduce the quality of the recommendations. The algorithm updates immediately on new information about a person's interests. Finally, the recommendations can be explained in an intuitive way as arising from a list of items the customer remembers purchasing.

In 2003: Amazon.com, Netflix, YouTube, and More

By the time we published in IEEE in 2003, item-based collaborative filtering was widely deployed across Amazon.com. The homepage prominently featured recommendations based on your past purchases and items browsed in the store. Search result pages recommended items related to your search. The shopping cart recommended other items to add to your cart, perhaps impulse buys to bundle in at the last minute, or perhaps complements to what you were already considering. At the end of your order, more recommendations appeared, suggesting items to order later. Using e-mails, browse pages, product detail pages, and more, many pages on Amazon.com had at least some recommendations content, starting to approach a store for every customer.

Others have reported using the algorithm, too. In 2010, YouTube reported using it for recommending videos.[3] Many open source and third-party vendors included the algorithm, and it showed up widely in online retail, travel, news, advertising, and more. In the years following, the recommendations were used so extensively by Amazon.com that a Microsoft Research report estimated 30 percent of Amazon.com's page

G.D. Linden, J.A. Jacobi, and E.A. Benson, Collaborative Recommendations Using Item-to-Item Similarity Mappings, US Patent 6,266,649, to Amazon.com, Patent and Trademark Office, 2001 (filed **1998**).

https://www.computer.org/csdl/mags/ic/2017/03/mic2017030012.html

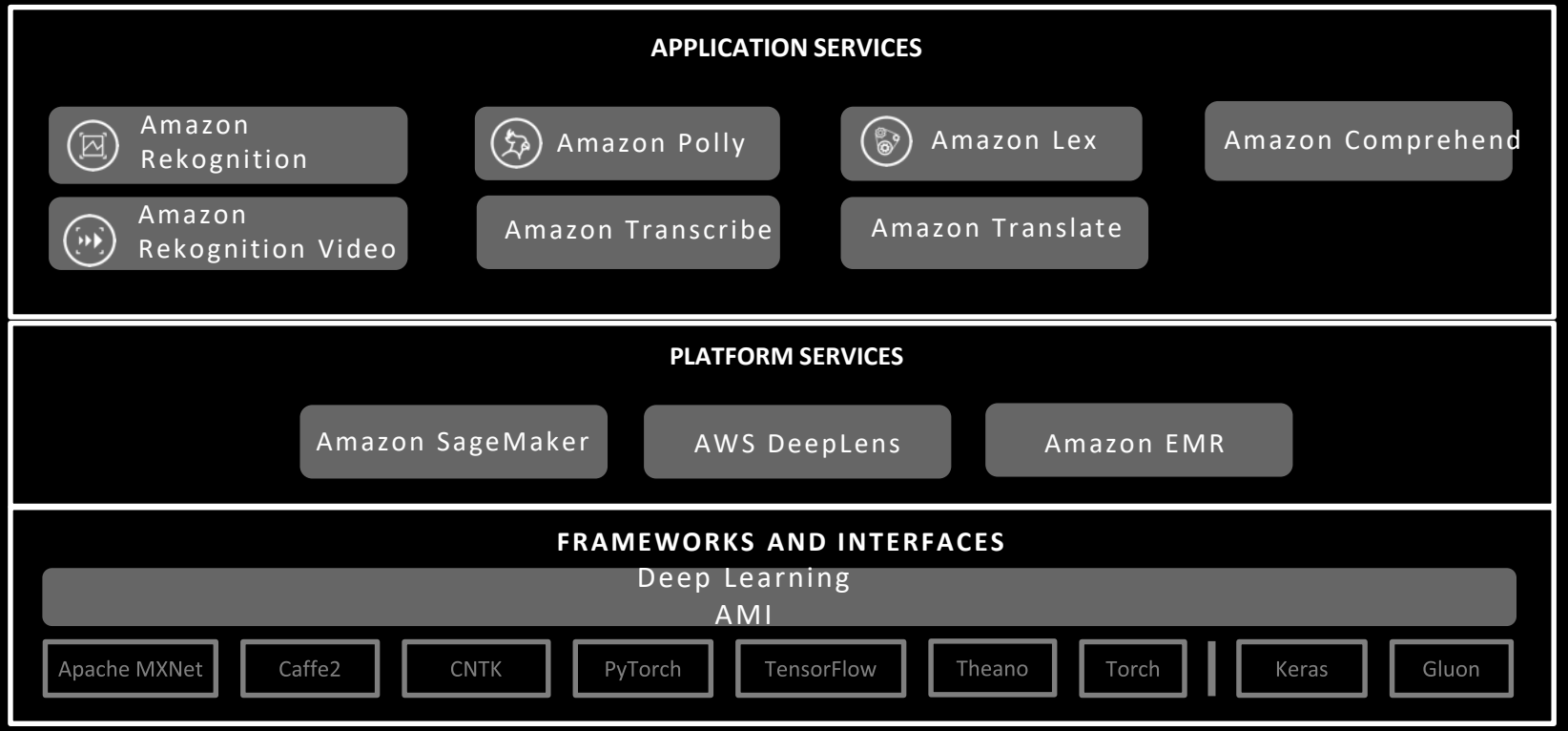

1C↑
2C↑
3C↑
4C↑
1B↑
2B↑
3B↑
4B↑
1A↑
2A↑
3A↑
4A↑
750 lbs. BASE CAPACITY
DO NOT EXCEED LIMIT
2446
amazonrobotics

amazon echo


INTRODUCING
amazon go


amazon
PrimeAir
amazon
PrimeAir

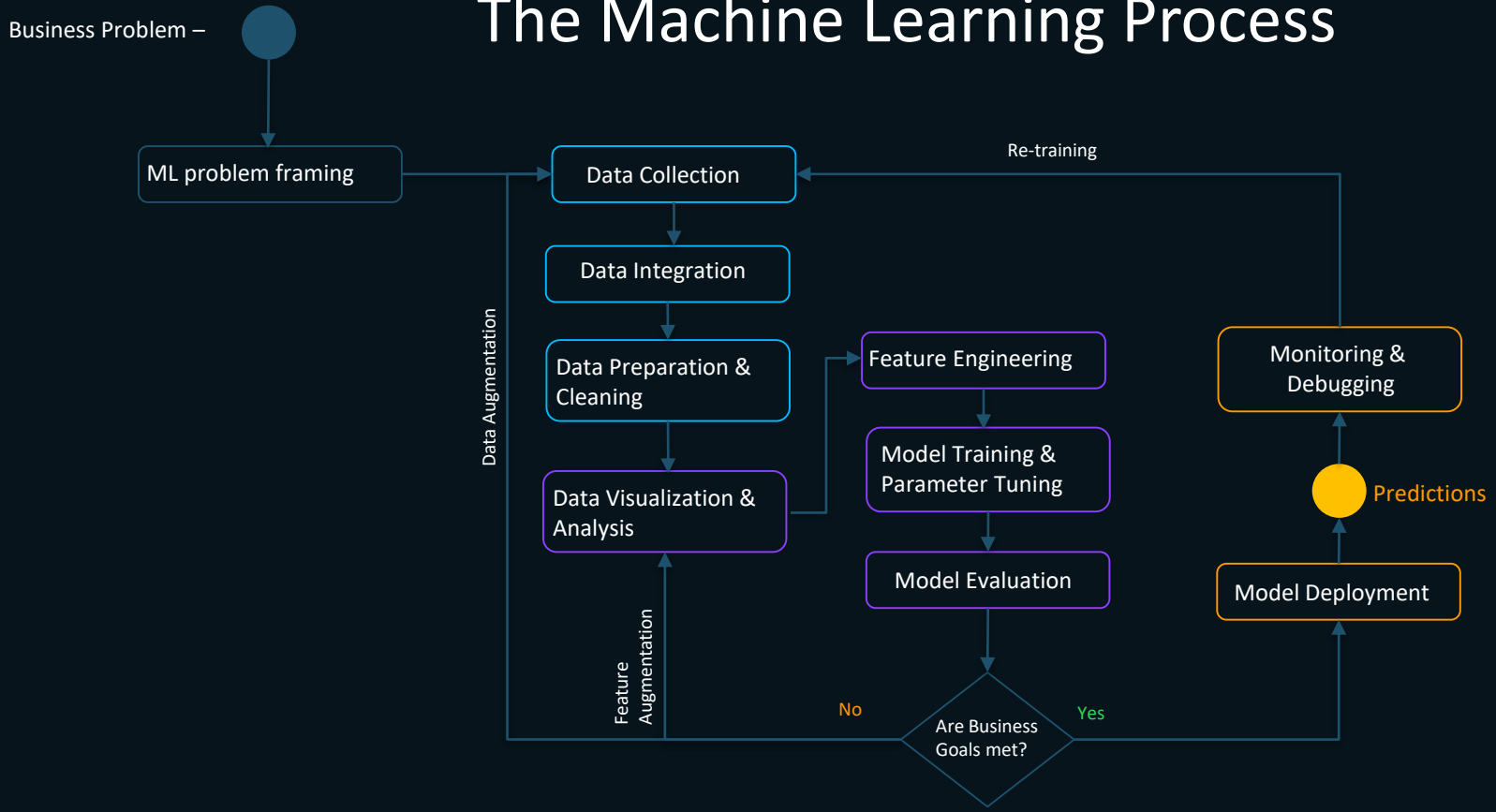
Democratization of AI



More AI/ML is built on AWS than anywhere else

More AI/ML is built on AWS than anywhere else

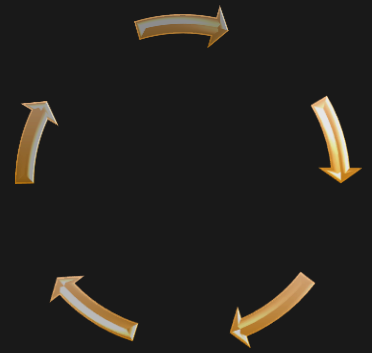
The Machine Learning Process



Amazon SageMaker

Build

Pre-built notebook instances



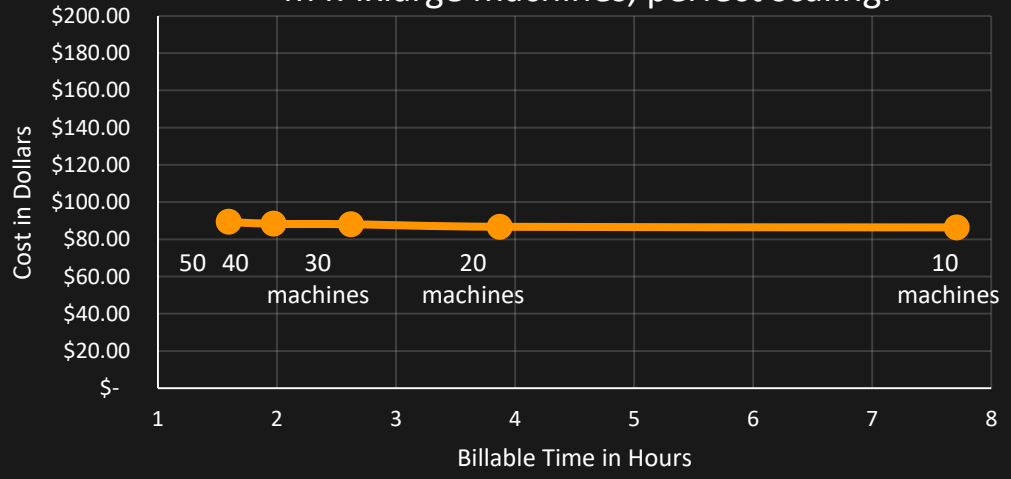
Highly-optimized machine learning algorithms

Factorization Machines

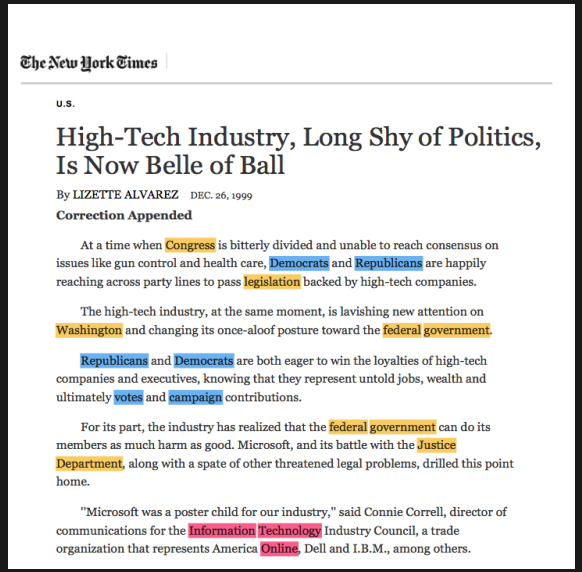
$$\tilde{y} = w_0 + \langle w_1, x \rangle + \sum_{i,j>i} x_i x_j \cdot \langle v_i, v_j \rangle$$

	Log_loss	F1 Score	**Seconds**
SageMaker	0.494	0.277	820
Other (10 Iter)	0.516	0.190	650
Other (20 Iter)	0.507	0.254	1300
Other (50 Iter)	0.481	0.313	3250

Click Prediction 1 TB advertising dataset, m4.4xlarge machines, perfect scaling.



Spectral LDA



The New York Times

U.S.

High-Tech Industry, Long Shy of Politics, Is Now Belle of Ball

By LIZETTE ALVAREZ DEC. 26, 1999

Correction Appended

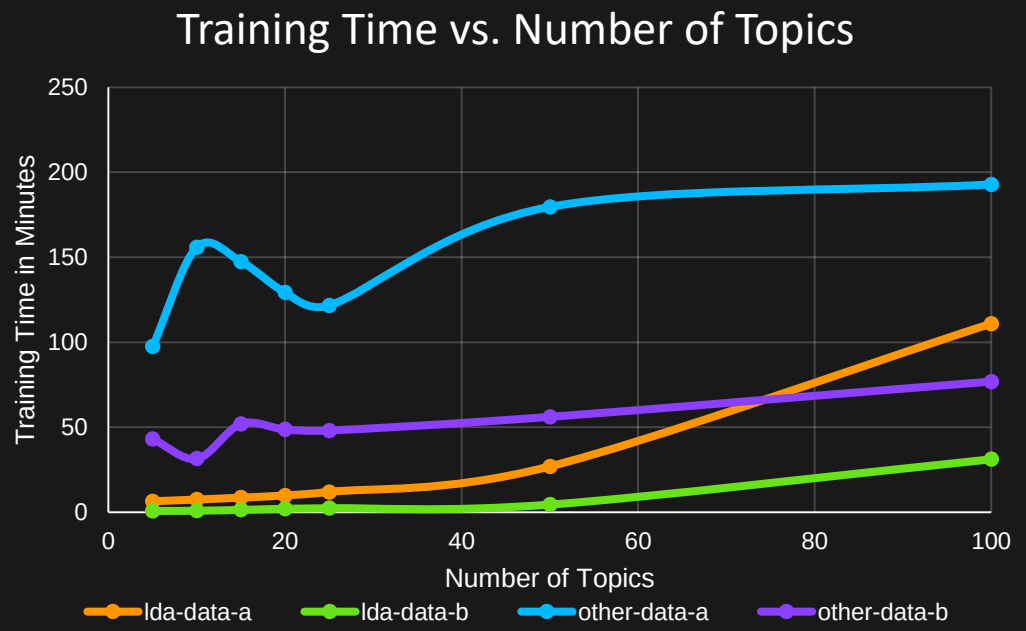
At a time when Congress is bitterly divided and unable to reach consensus on issues like gun control and health care, Democrats and Republicans are happily reaching across party lines to pass legislation backed by high-tech companies.

The high-tech industry, at the same moment, is lavishing new attention on Washington and changing its once-aloof posture toward the federal government.

Republicans and Democrats are both eager to win the loyalties of high-tech companies and executives, knowing that they represent untold jobs, wealth and ultimately votes and campaign contributions.

For its part, the industry has realized that the federal government can do its members as much harm as good. Microsoft, and its battle with the Justice Department, along with a spate of other threatened legal problems, drilled this point home.

"Microsoft was a poster child for our industry," said Connie Correll, director of communications for the Information Technology Industry Council, a trade organization that represents America Online, Dell and I.B.M., among others.



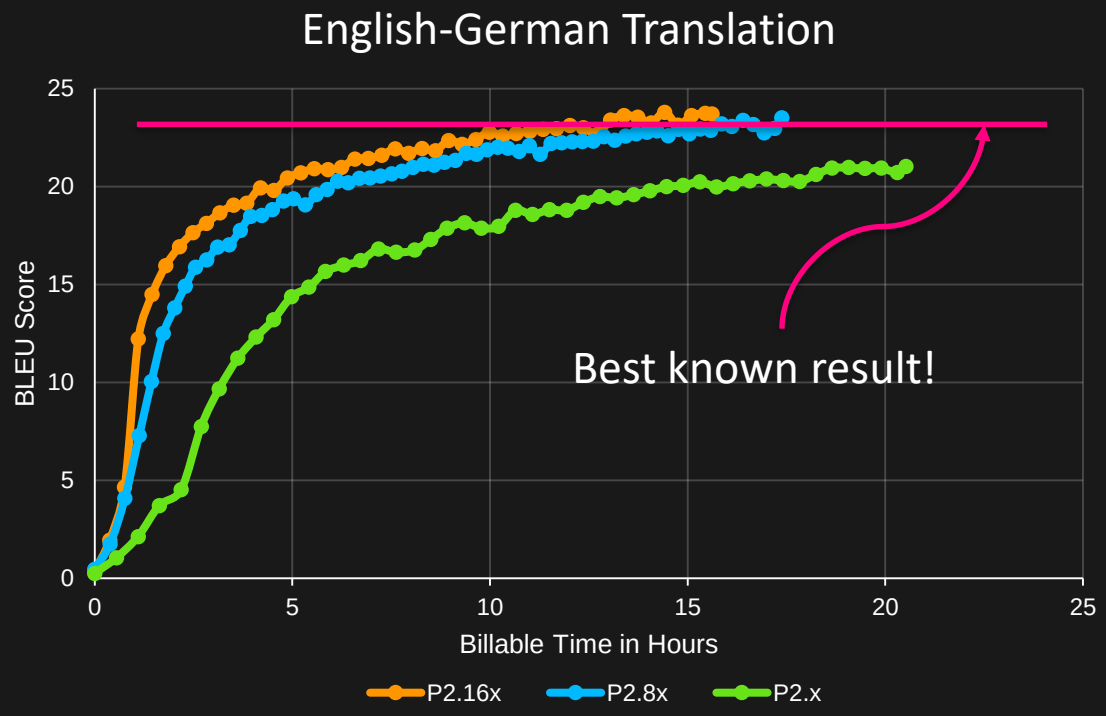
Sequence to Sequence

Based on Sockeye and Apache MXNet.

Supports multi-GPU training.

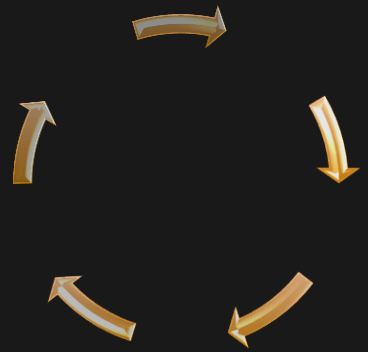
Can be used for Neural Machine Translation.

Supports both RNN/CNN as encoder/decoder



https://arxiv.org/abs/1712.05690
https://github.com/awslabs/sockeye

Amazon SageMaker



Pre-built notebook instances

Build

Highly-optimized machine learning algorithms

scikit learn

mxnet

GLUON

Train

One-click training for ML, DL, and custom algorithms

TensorFlow

PYTORCH

Easier training with hyperparameter optimization

aws

Amazon SageMaker

Build

Pre-built notebook instances

Highly-optimized machine learning algorithms

scikit learn

mxnet

GLUON

Train

One-click training for ML, DL, and custom algorithms

TensorFlow

PYTORCH

Easier training with hyperparameter optimization

Deploy

Fully-managed hosting at scale

Deployment without engineering effort

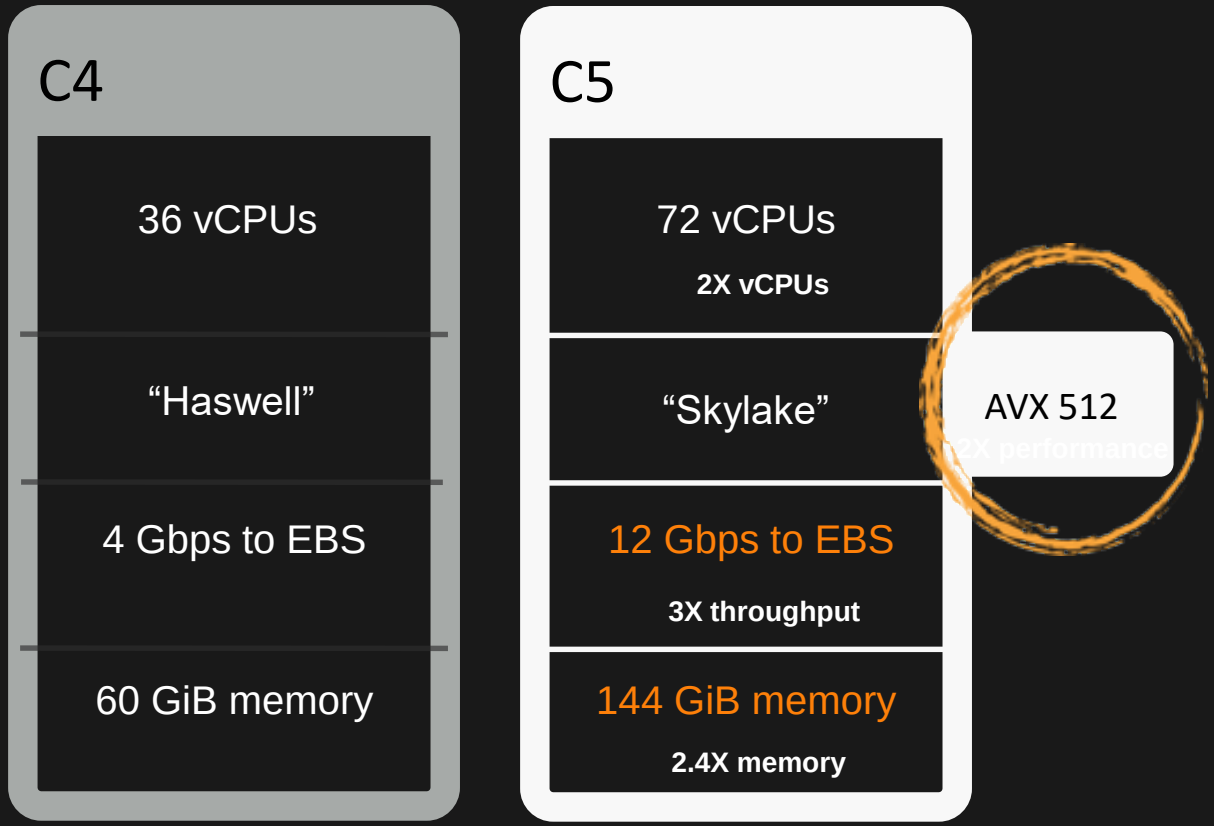
aws

Infrastructure – Amazon EC2 P3 Instances

The fastest, most powerful GPU instances in the cloud

- Up to eight NVIDIA Tesla V100 GPUs
- 1 PetaFLOPs of computational performance *– 14x better than P2*
- 300 GB/s GPU-to-GPU communication (NVLink) *– 9X better than P2*
- 16GB GPU memory with 900 GB/sec peak GPU memory bandwidth

Infrastructure – Amazon EC2 C5



DEMO

Resources

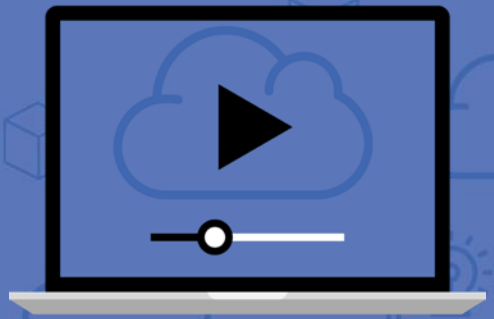
https://aws.amazon.com/machine-learning
https://aws.amazon.com/blogs/ai

https://aws.amazon.com/sagemaker
https://github.com/aws/sagemaker-python-sdk
https://github.com/aws/sagemaker-spark
https://github.com/awslabs/amazon-sagemaker-examples

An overview of Amazon SageMaker
https://www.youtube.com/watch?v=ym7NEYEx9x4

https://medium.com/@julsimon

Thank you!



Julien Simon, AI Evangelist, EMEA
@julsimon