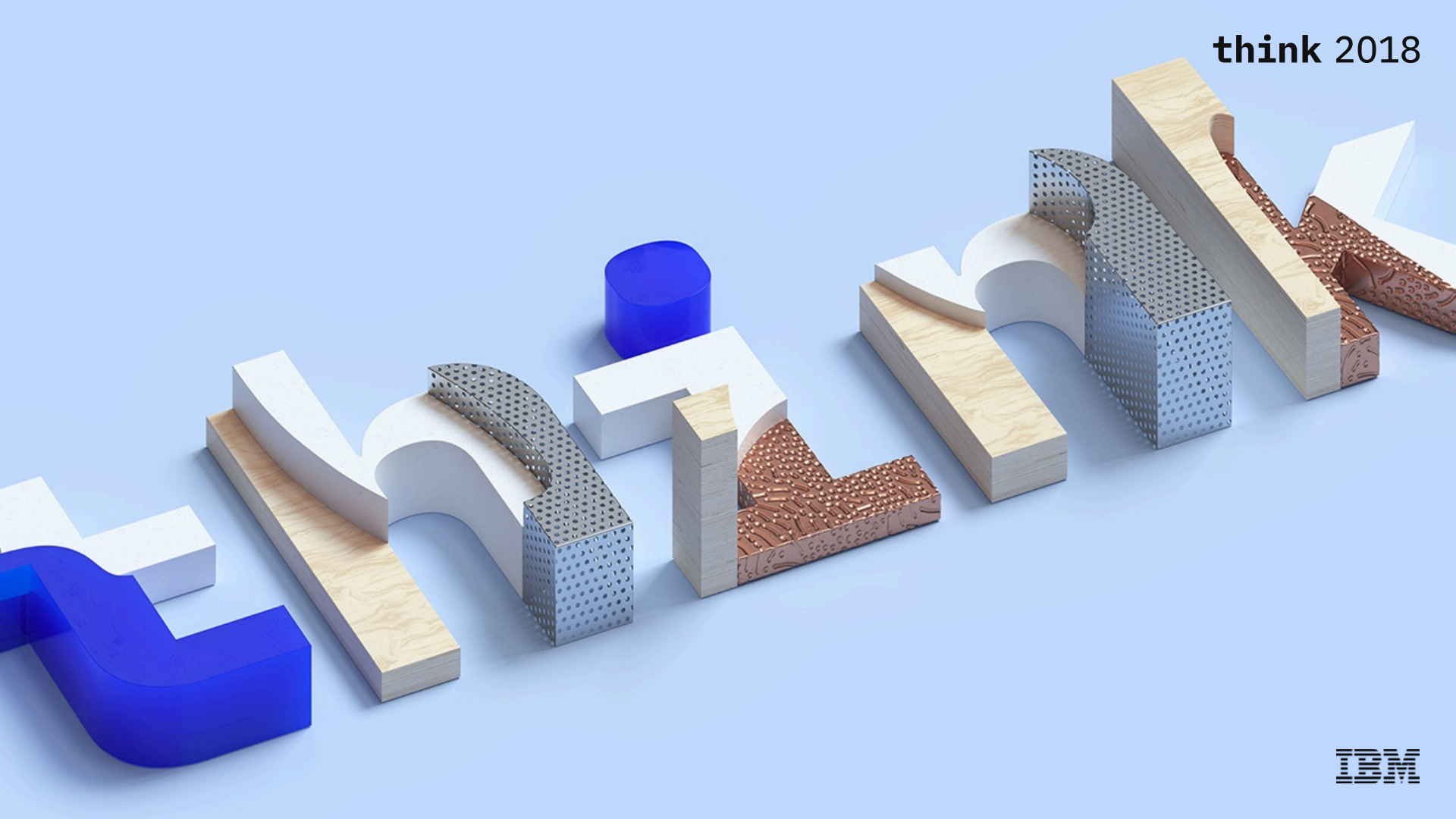

think 2018
IBM

Demystifying AI, Machine Learning & Deep Learning

Session 7918

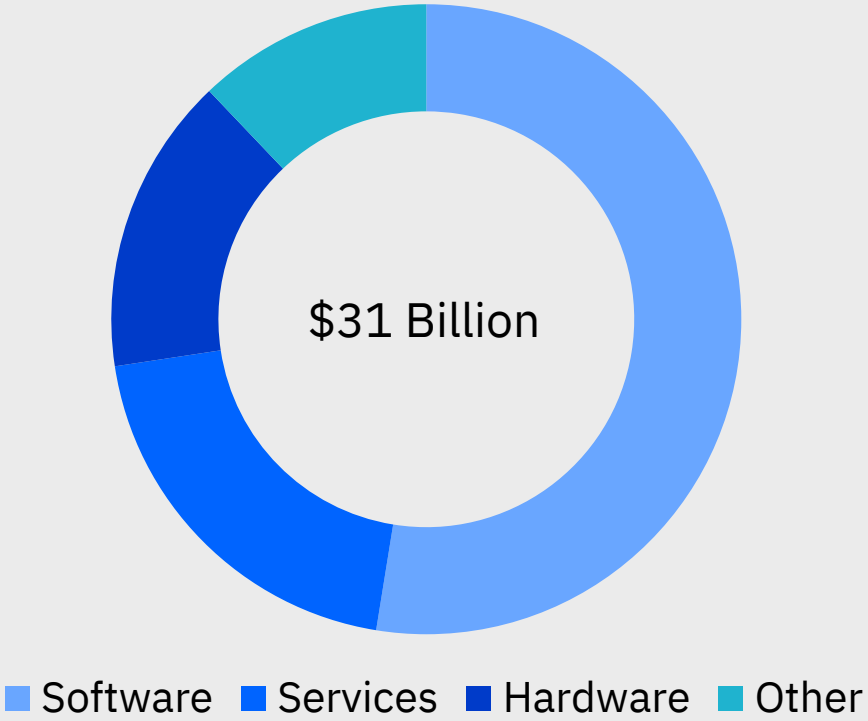
Sumit Gupta
VP, AI & Machine Learning
IBM Cognitive Systems

think 2018

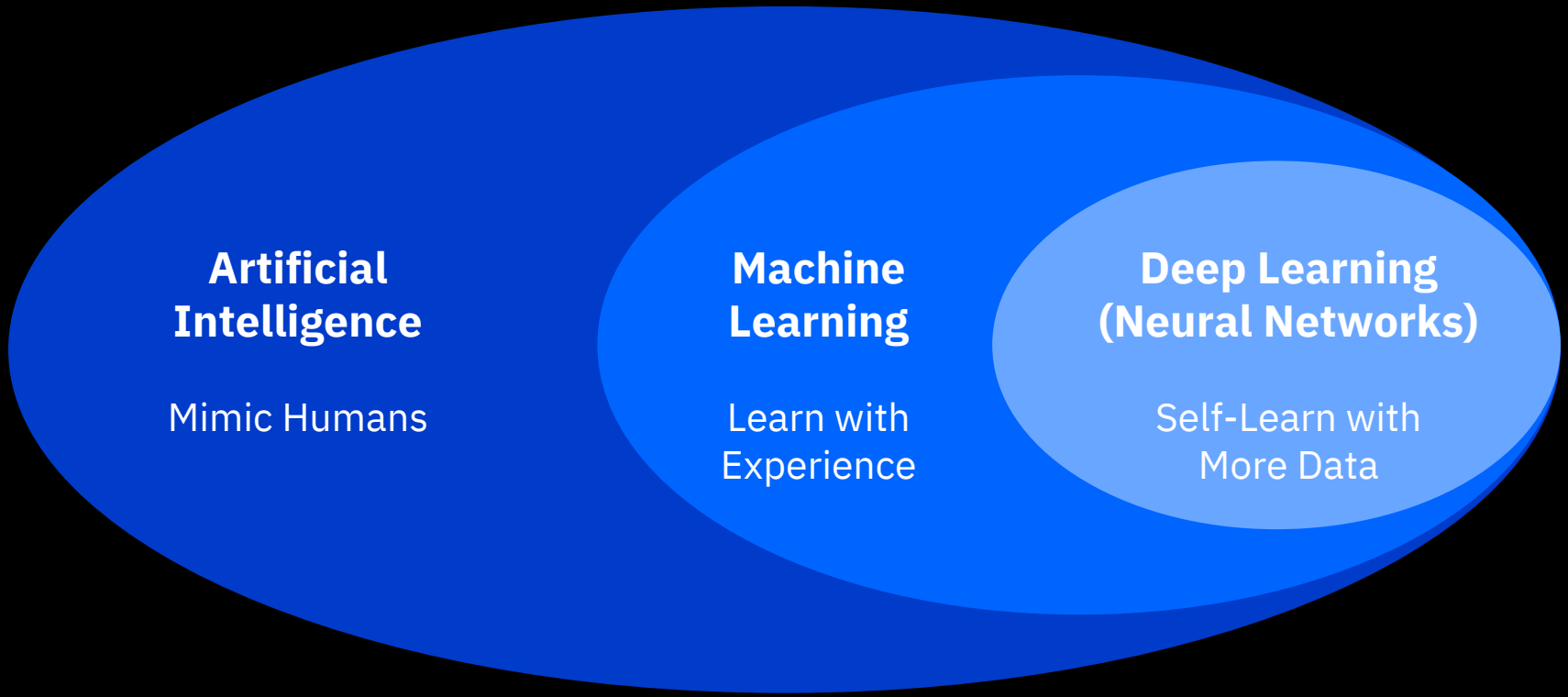
Cognitive IT Spend will Grow 3x in 2 Years

IDC Market Data

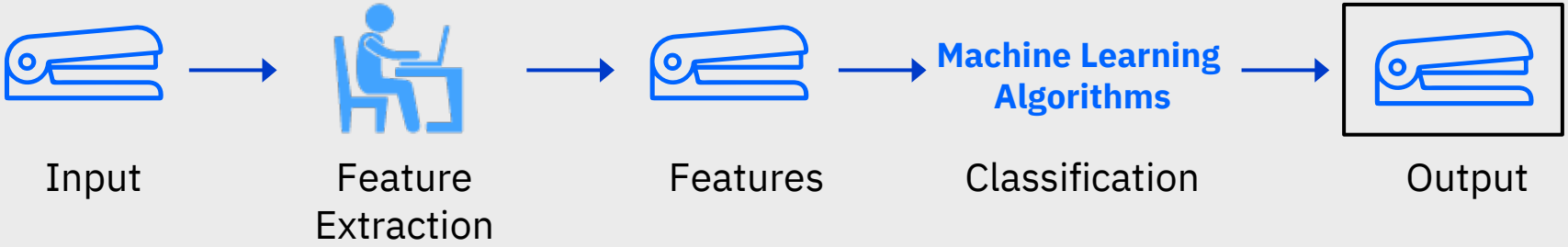
Cognitive Market Spend (2019)



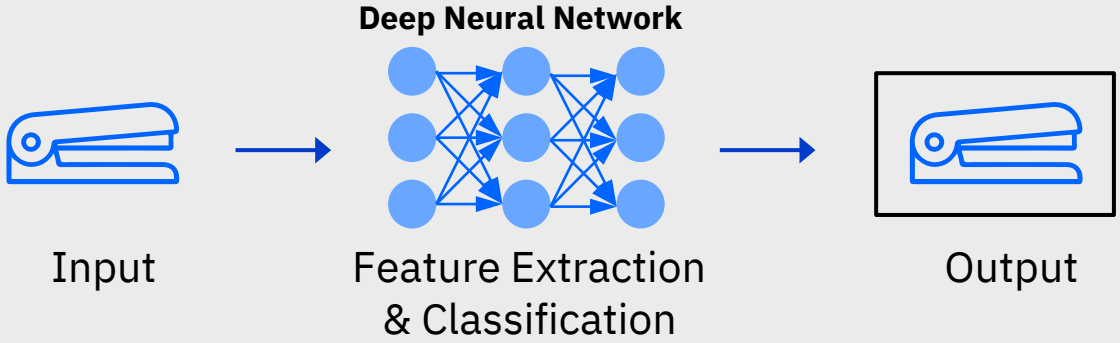
Teaching a Computer to Recognize a Bicycle


Artificial Intelligence
Mimic Humans
Machine Learning
Learn with Experience
Deep Learning (Neural Networks)
Self-Learn with More Data

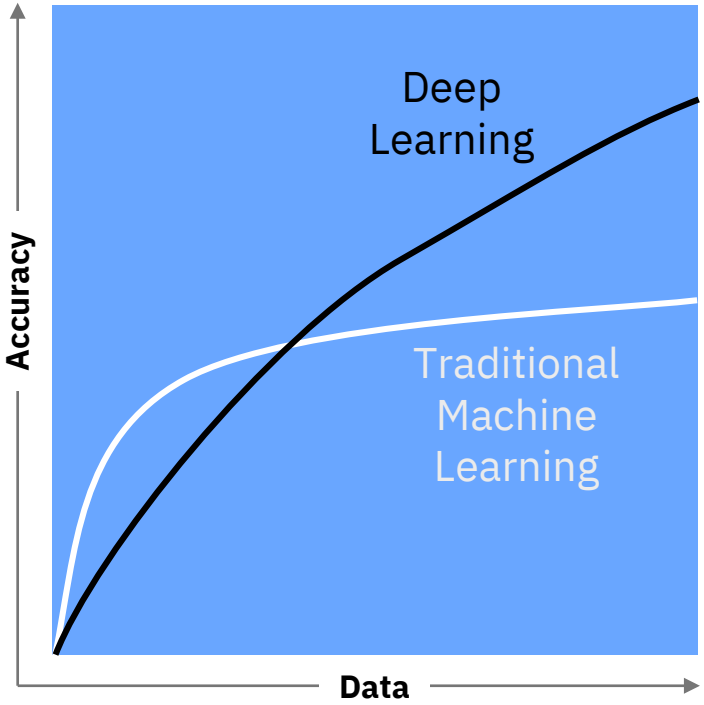
Machine Learning



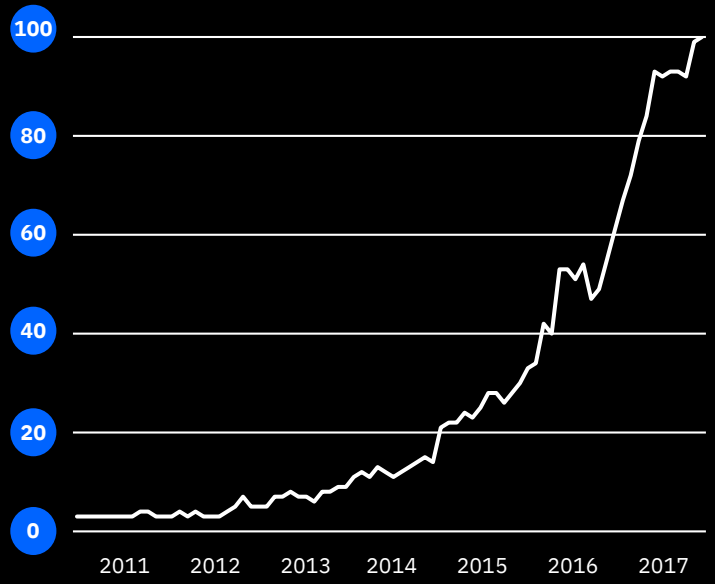
Deep Learning



Deep Learning Has Revolutionized Machine Learning



of Searches for Deep Learning from 2011 to 2017



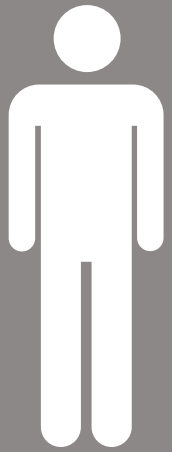
Source: Google Trends. Search term “Deep Learning”

2011

26% Errors

Machine Learning Based

Humans



5% Error

2016

3% Errors

Deep Learning Based

AI Enterprise Use Cases

Financial Services

- Fraud detection, Credit Risk Analysis, Customer cross-sell
- Predict end of day deposit levels required
- Vision based: ID, Fraud, Insurance claims

Retailers

- Recommendation engines using POS data, E-tail image analytics
- Computer vision for shopping carts, manufacturing
- Customer Insights

Customer Service

- Understand speech, language, Q&A, enable smarter search
- Sentiment & Tone Analysis

Transportation & Industrial IoT

- Computer vision for self-driving vehicles
- Industrial / Inventory Inspection
- Predictive Maintenance

AUTOMOTIVE

ADAS, Maintenance

COMMUNICATIONS

Location-based advertising

CONSUMER GOODS

Sentiment analysis

FINANCIAL SERVICES

Risk analysis
Fraud detection

RESEARCH

Physics Modeling

MANUFACTURING

Line inspection,
Defect analysis

LIFE SCIENCES

Sequence Analysis,
Radiology

MEDIA/ENTERTAIN

Advertising effectiveness

CUSTOMER SERVICE

Chatbots, Helpdesk
Automated Expenses

HEALTH CARE

Patient sensors, monitoring, EHRs

OIL & GAS

Exploration,
sensor analysis

RETAIL

Recommendation Engines,

Precision Marketing

TRANSPORTATION

Optimal traffic flows,
Route planning

UTILITIES

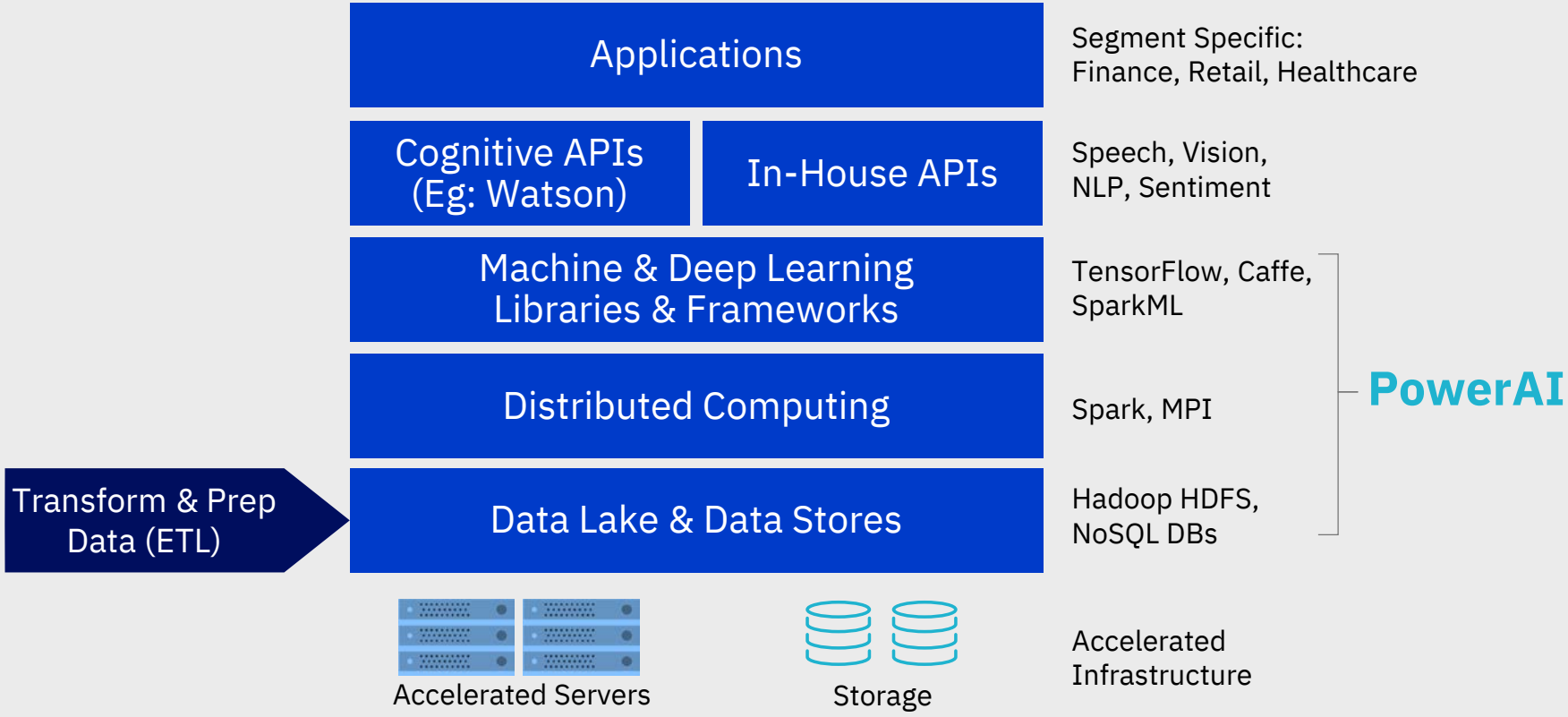
Smart Meter analysis,
Capacity planning

LAW & DEFENSE

Threat analysis –
social media monitoring

Richard Liu
VP, Quantitative Analytics
Wells Fargo

AI Infrastructure Stack



PowerAI

Integrated & Supported AI Platform
Higher Productivity for Data Scientists
Enable non-Data Scientists to use AI

Developer Ease-of-Use Tools

Open Source Frameworks:
Supported Distribution

TensorFlow™ Caffe torch Chainer

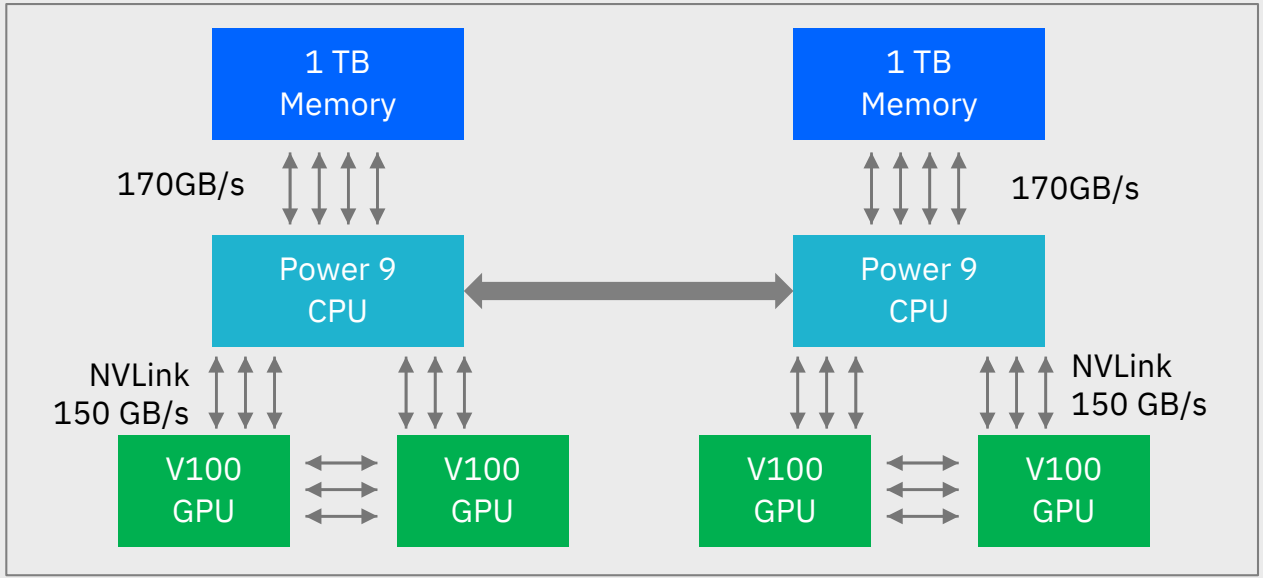
Faster Training Times via
HW & SW Performance Optimizations

5x Faster Data Communication with Unique CPU-GPU NVLink High-Speed Connection

Store Large Models in System Memory

Fast Transfer via NVLink

Operate on One Layer at a Time



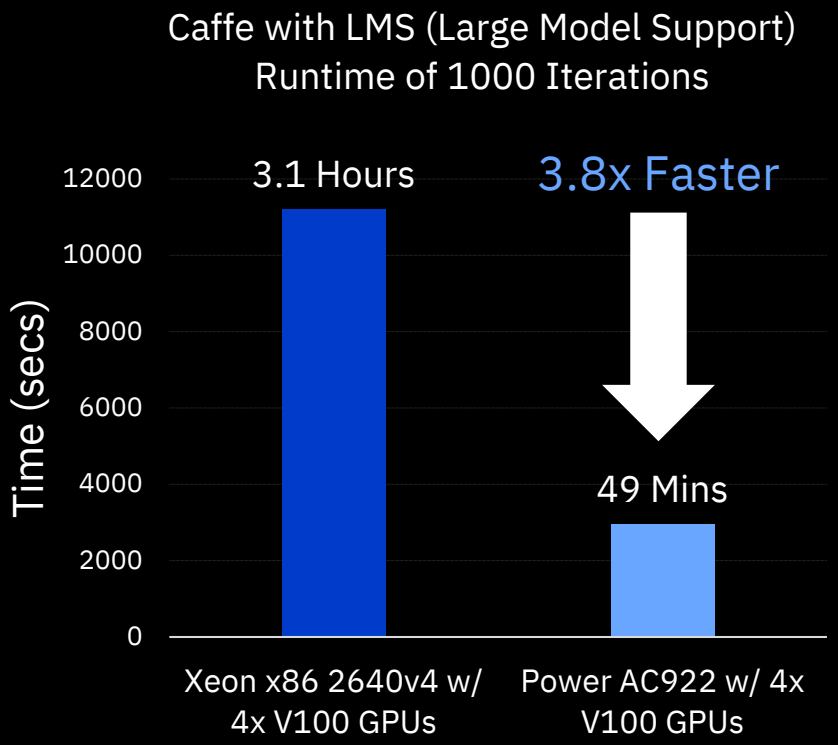
IBM AC922 Power System

Deep Learning Server (4-GPU Config)

Large AI Models Train ~4 Times Faster

POWER9 Servers with NVLink to GPUs
vs
x86 Servers with PCIe to GPUs

Detailed Benchmark Information in Back



GoogleNet model on Enlarged ImageNet Dataset (2240x2240)

Distributed Deep Learning (DDL)

Deep learning training takes days to weeks

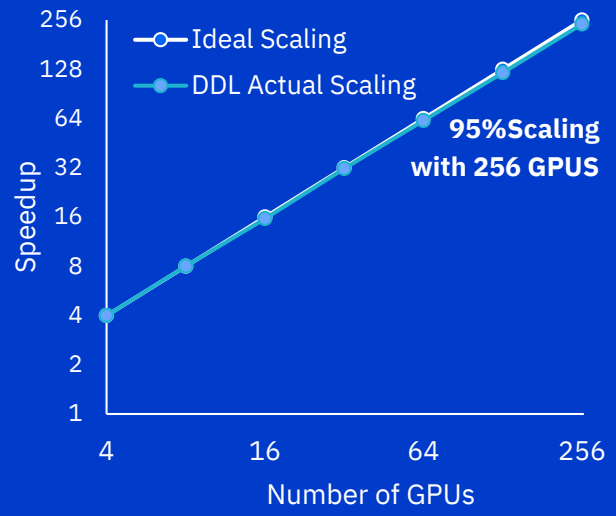
Limited scaling to multiple x86 servers

PowerAI with DDL enables scaling to 100s of servers

16 Days Down to 7 Hours
58x Faster

ResNet-101, ImageNet-22K

Near Ideal Scaling to 256 GPUs

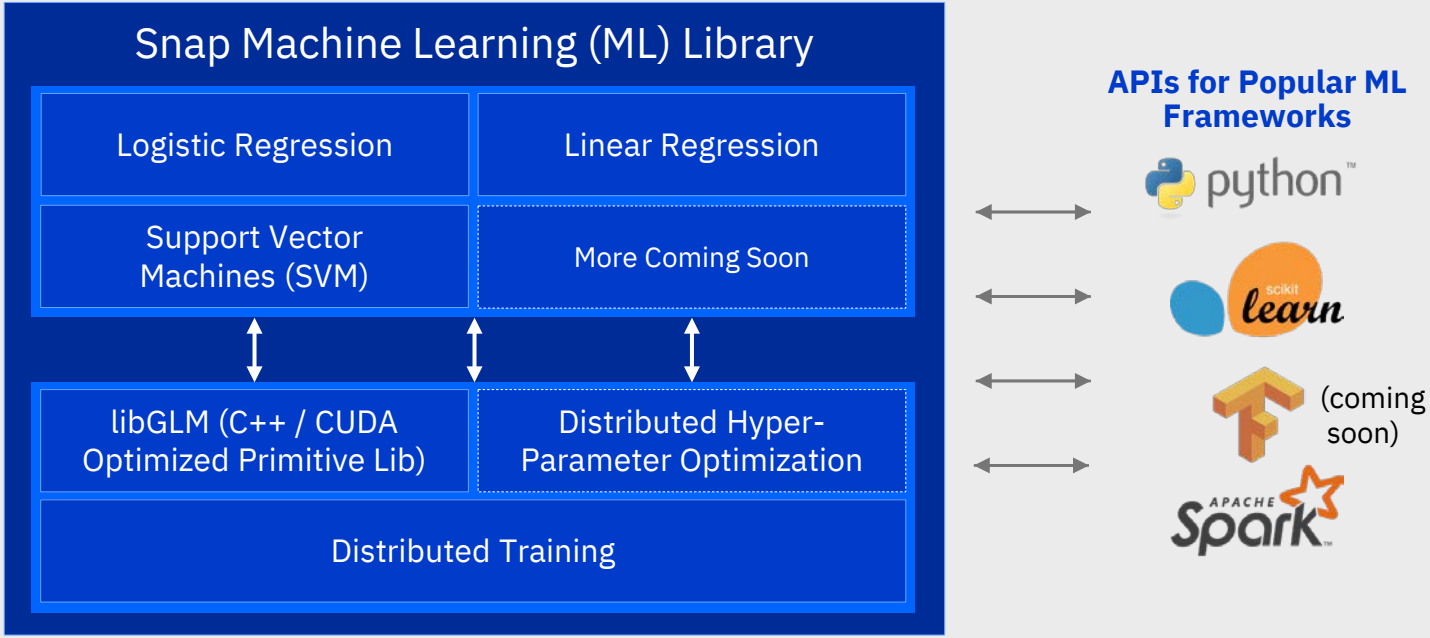


ResNet-50, ImageNet-1K

Caffe with PowerAI DDL, Running on Minsky (S822Lc) Power System

Snap ML

Distributed GPU-Accelerated Machine Learning Library



Snap ML: Training Time Goes From An Hour to Minutes

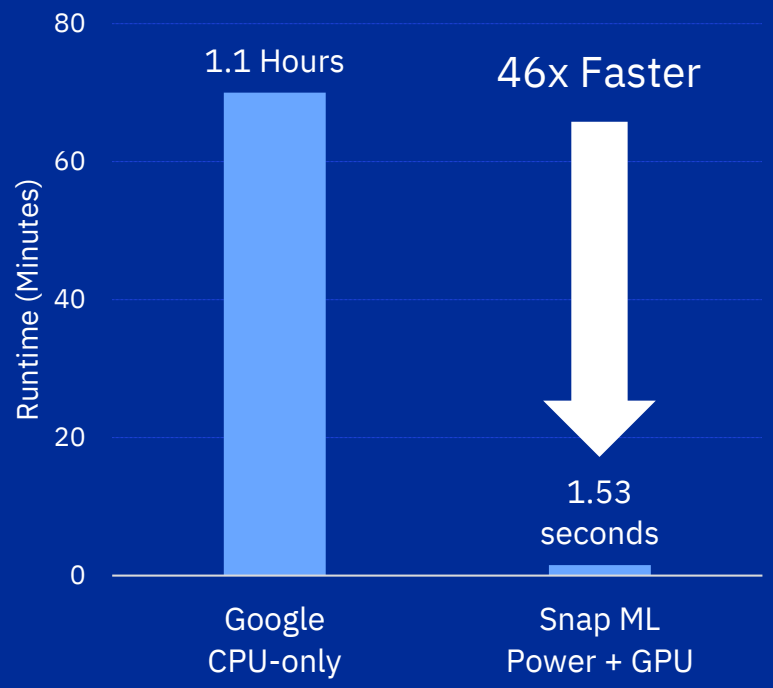
46x faster than previous record set by Google

Workload: Click-through rate prediction for advertising

Logistic Regression Classifier in Snap ML using GPUs vs TensorFlow using CPU-only

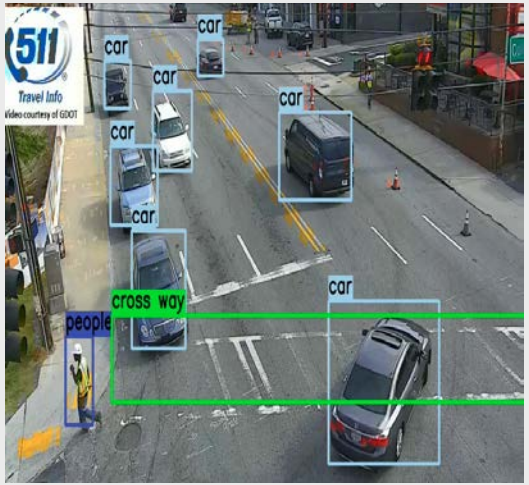
Dataset: Criteo Terabyte Click Logs (http://labs.criteo.com/2013/12/download-terabyte-click-logs/)
4 billion training examples, 1 million features
Model: Logistic Regression: TensorFlow vs Snap ML
Test LogLoss: 0.1293 (Google using Tensorflow), 0.1292 (Snap ML)
Platform: 89 CPU-only machines in Google using Tensorflow versus 4 AC922 servers (each 2 Power9 CPUs + 4 V100 GPUs) for Snap ML
Google data from this Google blog

Logistic Regression in Snap ML (with GPUs) vs TensorFlow (CPU-only)

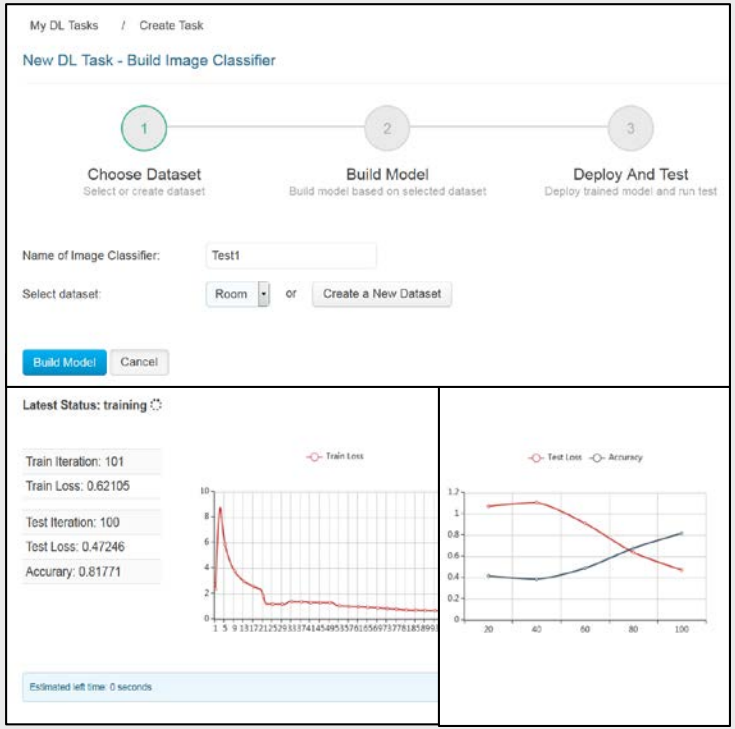


PowerAI Vision: ”Point-and-Click” AI for Images & Video

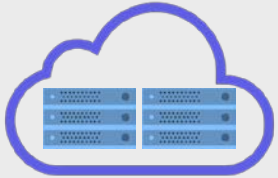
Label Image or Video Data



Auto-Train AI Model



Package & Deploy AI Model



Get Started Today with Machine & Deep Learning

IBM PowerAI

Build a Data Science Team
Your Developers Can Learn
http://cognitiveclass.ai

Identify a Low Hanging Use Case

Figure Out Data Strategy

Consider Pre-Built AI APIs

Hire Consulting Services

Get Started Today at
www.ibm.biz/poweraideveloper

Please visit one of the following Think Tanks to learn more.

12:30 PM – 12:50 PM
1:00 PM – 1:20 PM

Think Tank A
2422

Optimizing Accelerated Cognitive Systems with OpenCAPI

Think Tank B
7767

Introducing Deep Learning for Vision Analytics with IBM PowerAI Vision

Think Tank C
8060

Unlocking Shared Memory with Coherence

Think Tank D
8058

Deep Learning-Based AI for the Enterprise

Large Model Support Benchmark Details

- Hardware: Power AC922; 40 cores (2 x 20c chips), POWER9 with NVLink 2.0; 2.25 GHz, 1024 GB memory, 4xTesla V100 GPU Pegas 1.0. Competitive stack: 2x Xeon E5-2640 v4; 20 cores (2 x 10c chips) / 40 threads; Intel Xeon E5-2640 v4; 2.4 GHz; 1024 GB memory, 4xTesla V100 GPU, Ubuntu 16.04.
- Chainer: IBM Internal Measurements running 1000 iterations of Enlarged GoogleNet model on Enlarged Imagenet Dataset (2240x2240) .
    - Software: Chainverv3 /LMS/Out of Core with CUDA 9 / CuDNN7 with patches found at https://github.com/cupy/cupy/pull/694 and https://github.com/chainer/chainer/pull/3762
- Caffe Results: IBM Internal Measurements running 1000 iterations of Enlarged GoogleNet model (mini-batch size=5) on Enlarged Imagenet Dataset (2240x2240) .
    - Software: IBM Caffe with LMS Source code: https://github.ibm.com/TUNG/trlcaffe/tree/1.0-ibm-blc-bm-fix-hang+-p9collateral based on the branch "1.0-ibm-blc-bm-fix-hang+" (base for PowerAI R4) and a PR#5972 from BVLC/Caffe (for supporting cudnn7).

Notices and disclaimers

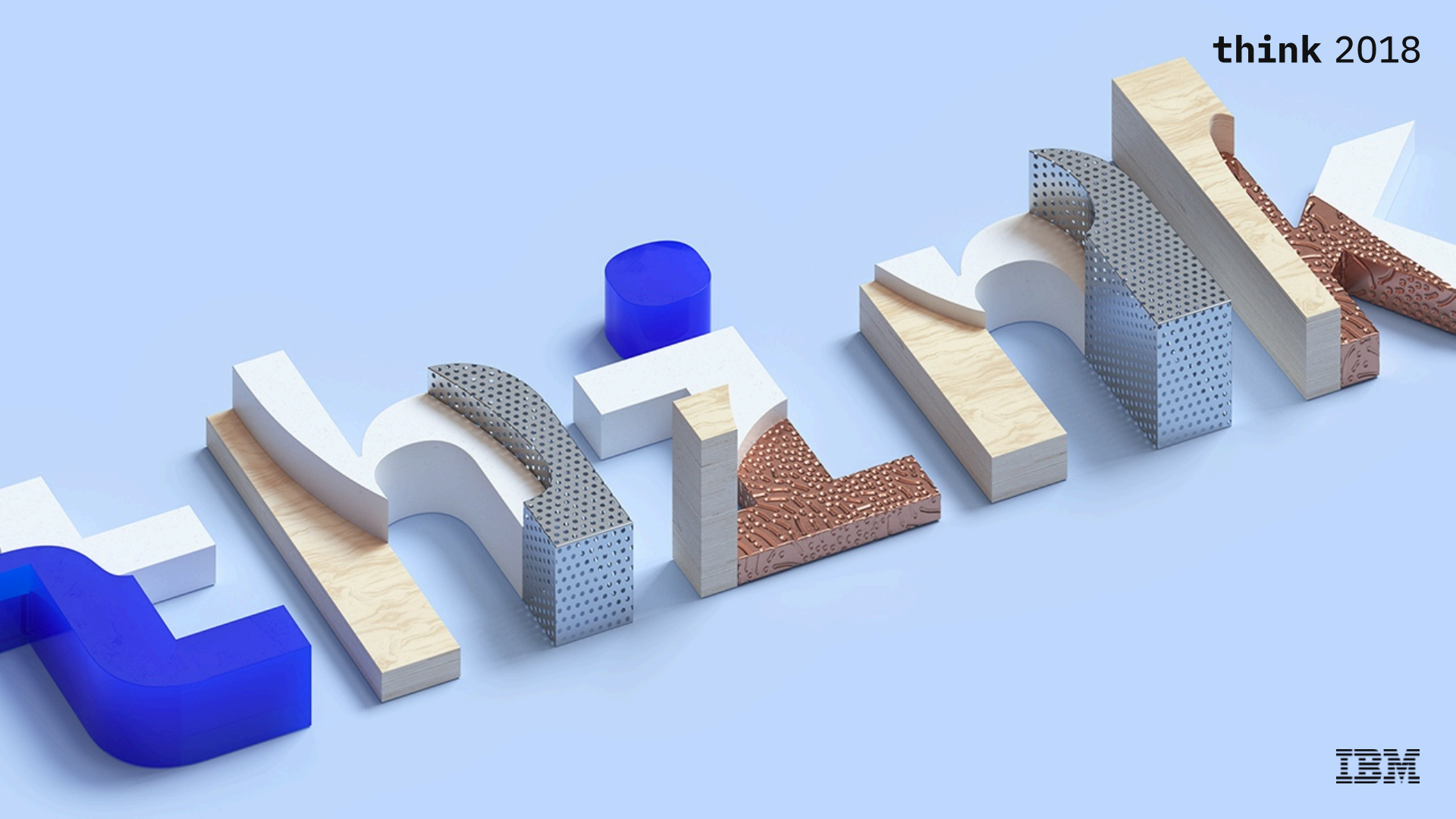
© 2018 International Business Machines Corporation. No part of this document may be reproduced or transmitted in any form without written permission from IBM.
U.S. Government Users Restricted Rights — use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM.
Information in these presentations (including information relating to products that have not yet been announced by IBM) has been reviewed for accuracy as of the date of initial publication and could include unintentional technical or typographical errors. IBM shall have no responsibility to update this information. **This document is distributed "as is" without any warranty, either express or implied. In no event, shall IBM be liable for any damage arising from the use of this information, including but not limited to, loss of data, business interruption, loss of profit or loss of opportunity.** IBM products and services are warranted per the terms and conditions of the agreements under which they are provided.
IBM products are manufactured from new parts or new and used parts. In some cases, a product may not be new and may have been previously installed. Regardless, our warranty terms apply."
Any statements regarding IBM's future direction, intent or product plans are subject to change or withdrawal without notice.

Performance data contained herein was generally obtained in a controlled, isolated environments. Customer examples are presented as illustrations of how those
customers have used IBM products and the results they may have achieved. Actual performance, cost, savings or other results in other operating environments may vary.
References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business.
Workshops, sessions and associated materials may have been prepared by independent session speakers, and do not necessarily reflect the views of IBM. All materials and discussions are provided for informational purposes only, and are neither intended to, nor shall constitute legal or other guidance or advice to any individual participant or their specific situation.
It is the customer's responsibility to insure its own compliance with legal requirements and to obtain advice of competent legal counsel as to the identification and interpretation of any relevant laws and regulatory requirements that may affect the customer's business and any actions the customer may need to take to comply with such laws. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the customer follows any law.

Notices and disclaimers continued

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products about this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products. IBM does not warrant the quality of any third-party products, or the ability of any such third-party products to interoperate with IBM's products. **IBM expressly disclaims all warranties, expressed or implied, including but not limited to, the implied warranties of merchantability and fitness for a purpose.** The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents, copyrights, trademarks or other intellectual property right.

IBM, the IBM logo, ibm.com and [names of other referenced IBM products and services used in the presentation] are trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at: www.ibm.com/legal/copytrade.shtml.


think 2018
IBM