DEEP LEARNING AND BEYOND

Will Ramey, Sr. Director, Global Head of Developer Programs, NVIDIA Corporation
ACCELERATED DATA SCIENCE

DATA ANALYTICS
Extracting insights from big data

MACHINE LEARNING
Learning from examples in the data

DEEP LEARNING
Automating feature engineering
GPU-ACCELERATED DATA SCIENCE

Use Cases in Every Industry

**CONSUMER INTERNET**
- Ad Personalization
- Click Through Rate Optimization
- Churn Reduction

**FINANCIAL SERVICES**
- Claim Fraud
- Customer Service Chatbots/Routing
- Risk Evaluation

**HEALTHCARE**
- Improve Clinical Care
- Drive Operational Efficiency
- Speed Up Drug Discovery

**RETAIL**
- Supply Chain & Inventory Management
- Price Management / Markdown Optimization
- Promotion Prioritization And Ad Targeting

**OIL & GAS**
- Sensor Data Tag Mapping
- Anomaly Detection
- Robust Fault Prediction

**MANUFACTURING**
- Remaining Useful Life Estimation
- Failure Prediction
- Demand Forecasting

**TELECOM**
- Detect Network/Security Anomalies
- Forecasting Network Performance
- Network Resource Optimization (SON)

**AUTOMOTIVE**
- Personalization & Intelligent Customer Interactions
- Connected Vehicle Predictive Maintenance
- Forecasting, Demand, & Capacity Planning

**ADDITIONAL USE CASES**
- **CONSUMER INTERNET**
  - **CHURN REDUCTION**
  - **CLAIM FRAUD**
  - **CUSTOMER SERVICE CHATBOTS/Routing**
  - **RISK EVALUATION**

- **FINANCIAL SERVICES**
  - **CLAIM FRAUD**
  - **CUSTOMER SERVICE CHATBOTS/Routing**
  - **RISK EVALUATION**

- **HEALTHCARE**
  - **IMPROVE CLINICAL CARE**
  - **DRIVE OPERATIONAL EFFICIENCY**
  - **SPEED UP DRUG DISCOVERY**

- **RETAIL**
  - **SUPPLY CHAIN & INVENTORY MANAGEMENT**
  - **PRICE MANAGEMENT / MARKDOWN OPTIMIZATION**
  - **PROMOTION PRIORITIZATION AND AD TARGETING**

- **OIL & GAS**
  - **SENSOR DATA TAG MAPPING**
  - **ANOMALY DETECTION**
  - **ROBUST FAULT PREDICTION**

- **MANUFACTURING**
  - **REMAINING USEFUL LIFE ESTIMATION**
  - **FAILURE PREDICTION**
  - **DEMAND FORECASTING**

- **TELECOM**
  - **DETECT NETWORK/SECURITY ANOMALIES**
  - **FORECASTING NETWORK PERFORMANCE**
  - **NETWORK RESOURCE OPTIMIZATION (SON)**

- **AUTOMOTIVE**
  - **PERSONALIZATION & INTELLIGENT CUSTOMER INTERACTIONS**
  - **CONNECTED VEHICLE PREDICTIVE MAINTENANCE**
  - **FORECASTING, DEMAND, & CAPACITY PLANNING**
BEYOND DEEP LEARNING
Opportunities to Accelerate Data Science

ARTIFICIAL INTELLIGENCE

Deep Learning

Machine Learning (Regressions, Decision Trees, Graph)

Analytics

Dense Data

Tabular/Sparse Data

MACHINE LEARNING / DATA ANALYTICS

2.2 exabytes (2.2B GB) of data created daily - McKinsey

$166B in 2018 revenues for big data and business analytics - IDC
RAPIDS DATA SCIENCE PLATFORM

GPU-accelerated Open Source Libraries

Data Preparation
- cuDF

Model Training
- cuML

Graph Analytics
- cuGRAPH

PYTHON

DEEP LEARNING FRAMEWORKS
- CUDNN

CUDA

APACHE ARROW in GPU Memory

DASK/SPARK

RAPIDS
- CUDF
- CUML
- CUGRAPH

Deep Learning Frameworks
- CUDNN
DL IS CRITICAL FOR INTERNET APPLICATIONS
Users Expect Intelligence in Services

Growing Use of Deep Learning at Google

Source: Jeff Dean, leads the Google Brain team, making machines intelligent. www.wsdm-conference.org/2016/slides/WSDM2016-Jeff-Dean.pdf
DEEP LEARNING IS SWEEPING ACROSS INDUSTRIES

Internet Services
- Image/Video Classification
- Speech Recognition
- Natural Language Processing

Medicine
- Cancer Cell Detection
- Diabetic Grading
- Drug Discovery

Media & Entertainment
- Video Captioning
- Content Based Search
- Real Time Translation

Security & Defense
- Face Recognition
- Video Surveillance
- Cyber Security

Autonomous Machines
- Pedestrian Detection
- Lane Tracking
- Recognize Traffic Signs
THE EXPANDING UNIVERSE OF MODERN AI

4,000+ AI START-UPS
$33B IN FUNDING

Source: Crunchbase & PitchBook
## WHAT PROBLEM ARE YOU SOLVING?
Defining the AI/DL task

<table>
<thead>
<tr>
<th>INPUTS</th>
<th>BUSINESS QUESTIONS</th>
<th>AI / DL TASK</th>
<th>EXAMPLE OUTPUTS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Is “it” present or not?</td>
<td>Detection</td>
<td>HEALTHCARE: Cancer Detection</td>
</tr>
<tr>
<td>Text Data</td>
<td>What type of thing is “it”?</td>
<td>Classification</td>
<td>HEALTHCARE: Image Classification</td>
</tr>
<tr>
<td>Images</td>
<td>To what extent is “it” present?</td>
<td>Segmentation</td>
<td>HEALTHCARE: Tumor Size / Shape Analysis</td>
</tr>
<tr>
<td>Video</td>
<td>What is the likely outcome?</td>
<td>Prediction</td>
<td>HEALTHCARE: Survivability Prediction</td>
</tr>
<tr>
<td>Audio</td>
<td>What will likely satisfy the objective?</td>
<td>Recommendations</td>
<td>HEALTHCARE: Therapy Recommendation</td>
</tr>
</tbody>
</table>
CAMBRIAN EXPLOSION

Convolutional Networks
- Encoder/Decoder
- ReLu
- BatchNorm
- Concat
- Dropout
- Pooling

Recurrent Networks
- LSTM
- GRU
- Beam Search
- WaveNet
- CTC
- Attention

Generative Adversarial Networks
- 3D-GAN
- MedGAN
- Conditional GAN
- Coupled GAN
- Speech Enhancement GAN

Reinforcement Learning
- DQN
- Simulation
- DDPG

New Species
- Mixture of Experts
- Neural Collaborative Filtering
- Block Sparse LSTM
DEEP LEARNING APPLICATION DEVELOPMENT
DEEP LEARNING APPLICATION DEVELOPMENT

Untrained Neural Network Model

TRAINING DATASET
DEEP LEARNING APPLICATION DEVELOPMENT

TRAINING
Learning a new capability from existing data

Untrained Neural Network Model

Deep Learning Framework

TRAINING DATASET

Deep Learning Application Development

Untrained Neural Network Model

Deep Learning Framework

TRAINING DATASET

Deep Learning Application Development
DEEP LEARNING APPLICATION DEVELOPMENT

TRAINING
Learning a new capability from existing data

Untrained Neural Network Model

Deep Learning Framework

TRAINING DATASET

Trained Model
New Capability

DEEP LEARNING APPLICATION DEVELOPMENT
DEEP LEARNING APPLICATION DEVELOPMENT

TRAINING
Learning a new capability from existing data

Untrained Neural Network Model

Deep Learning Framework

TRAINING DATASET

Trained Model New Capability

Trained Model Optimized for Performance
DEEP LEARNING APPLICATION DEVELOPMENT

TRAINING
Learning a new capability from existing data

Untrained Neural Network Model

Deep Learning Framework

TRAINING DATASET

Trained Model
New Capability

App or Service
Featuring Capability

NEW DATA

Trained Model Optimized for Performance

INFEERENCE
Applying this capability to new data
# CHALLENGES

<table>
<thead>
<tr>
<th>DEEP LEARNING NEEDS</th>
<th>WHY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Scientists</td>
<td>New computing model</td>
</tr>
<tr>
<td>Latest Algorithms</td>
<td>Rapid evolving</td>
</tr>
<tr>
<td>Fast Training</td>
<td>Impossible -&gt; Practical</td>
</tr>
<tr>
<td>Deployment Platforms</td>
<td>Must be available everywhere</td>
</tr>
</tbody>
</table>
ADVANCE YOUR DEEP LEARNING KNOWLEDGE AT GTC
Don’t miss the world’s most important event for GPU developers

WWW.GPUTECHCONF.COM
NVIDIA DEEP LEARNING INSTITUTE

Hands-on Training for Data Scientists and Software Engineers

Helping the world to solve challenging problems using AI and deep learning

On-site workshops and online courses presented by certified instructors

Covering complete workflows for proven application use cases

Fundamentals, Autonomous Vehicles, Finance, Healthcare, Video Analytics, IoT/Robotics, ...

www.nvidia.com/dli
Inception: Virtual Accelerator for AI Startups

**BENEFITS**

- AI Expertise
- Go-To-Market Support
- Technology Access

**MEMBERSHIP**

- 300
- 3600
- \(-10x\)

**IMPACT**

- Showcasing Innovation
- Creating a Global Community

www.nvidia.com/inception
# NVIDIA DEEP LEARNING SOFTWARE PLATFORM

## Training
- Training Data
- Data Management
- Training
- Model Assessment
- Trained Neural Network

## Inference
- Data center
- GRE + TensorRT
- Embedded
- JETPACK SDK
- Automotive
- DriveWorks SDK

## NVIDIA DEEP LEARNING SDK and CUDA

<table>
<thead>
<tr>
<th>cuDNN</th>
<th>NCCL</th>
<th>cuBLAS</th>
<th>cuSPARSE</th>
<th>TensorRT</th>
<th>DeepStream SDK</th>
</tr>
</thead>
<tbody>
<tr>
<td>![cuDNN Icon]</td>
<td>![NCCL Icon]</td>
<td>![cuBLAS Icon]</td>
<td>![cuSPARSE Icon]</td>
<td>![TensorRT Icon]</td>
<td>![DeepStream Icon]</td>
</tr>
</tbody>
</table>

developer.nvidia.com/deep-learning-software
NVIDIA GPU CLOUD
Simple Access to a Comprehensive Catalog of GPU-accelerated Software

Discover 30 GPU-Accelerated Containers
Deep learning, third-party managed HPC applications, NVIDIA HPC visualization tools, and partner applications

Innovate in Minutes, Not Weeks
Get up and running quickly and reduce complexity

Access from Anywhere
Use on PCs with NVIDIA Volta or Pascal™ architecture GPUs, NVIDIA DGX Systems, and supported cloud providers

ngc.nvidia.com
END-TO-END PRODUCT FAMILY

HPC / TRAINING

FULLY INTEGRATED AI SYSTEMS

DGX-1

DGX-2

DATA CENTER

Tesla P4/T4

Tesla V100

HSX1/HGX2

INFORMATION

CAR

EMBEDDED

DESKTOP

WORKSTATION

VIRTUAL WS

DATA CENTER

SERVER PLATFORM

TITAN

DGX Station

Virtual GPU

Tesla V100

Jetson AGX Xavier

Drive AGX Pegasus
# SOLUTIONS

<table>
<thead>
<tr>
<th>DEEP LEARNING NEEDS</th>
<th>SOLUTIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Scientists</td>
<td>Deep Learning Institute, GTC</td>
</tr>
<tr>
<td>Latest Algorithms</td>
<td>NGC, GPU Accelerated Frameworks, DL SDK</td>
</tr>
<tr>
<td>Fast Training</td>
<td>DGX, V100/T100, TITAN V</td>
</tr>
<tr>
<td>Deployment Platforms</td>
<td>TensorRT, V100/T4, Drive AGX, Jetson AGX</td>
</tr>
</tbody>
</table>
READY TO GET STARTED?

Project Checklist

What problem are you solving, what are the AI/DL tasks?

What data do you have/need, how is it labeled?

Which tools & environment will you use?

On what platform(s) will you train and deploy?
WHAT’S NEXT?

Work through DLI online courses  www.nvidia.com/dli
Review examples of AI in action news.developer.nvidia.com
Listen to the NVIDIA AI Podcast blogs.nvidia.com/ai-podcast
Register for GTC near you  www.nvidia.com/gtc

THANK YOU
NVIDIA DEEP LEARNING INSTITUTE

The NVIDIA Deep Learning Institute (DLI) offers hands-on training in artificial intelligence (AI) and accelerated computing to solve real-world problems. Start a self-paced online course at your convenience or take the course in an instructor-led workshop.

No matter how you learn best, DLI gives you access to a fully configured, GPU-accelerated workstation in the cloud, complete with software tools, neural networks, and datasets. All you need is a web browser and Internet connection to get started.

www.nvidia.com/dli