How to Improve Performance in Open-Source AI

Rachel Oberman, Intel AI Software Solutions Engineer

Agenda

- AI is Everywhere
  - Use Cases and Workflow Methodology
- Challenges with AI
- AI Performance Optimizations in Action (from Intel)

The Era of AI-based solutions is here

Agriculture

Achieving higher yields and efficiency

Energy

Increasing production and uptime

Education

Transforming learning

Government

Enhancing safety

Finance

Turning data into value

Healthcare

Revolutionizing patient outcomes

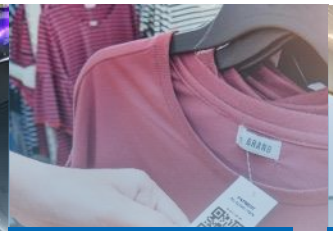
AI EVERYWHERE

Industrial

Empowering industry 4.0

Media

Creating thrilling experiences

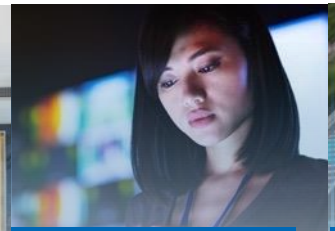


Retail

Modernizing shopping

Smart Home

Enabling homes to see, hear & respond



Telecom

Driving network efficiency

Transport

Fueling automated driving

Arts Enriched by AI – Neural Style Transfer

Many Approaches to Analytics & AI

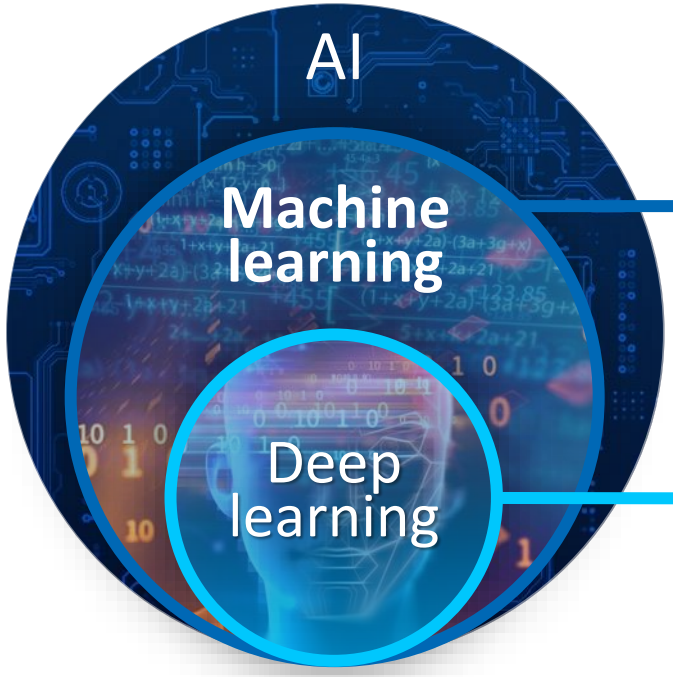
No one size fits all

Supervised Learning

Unsupervised Learning

Semi-Supervised Learning

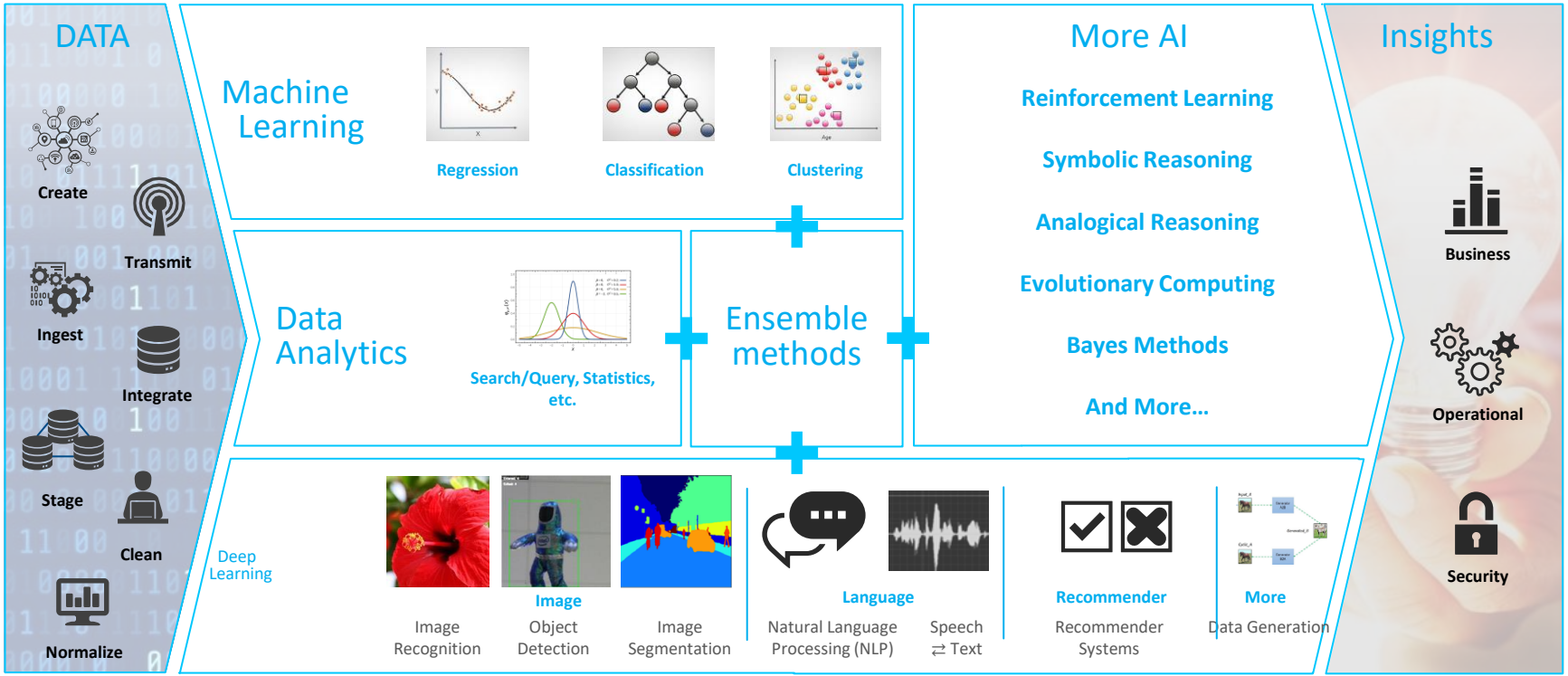
Reinforcement Learning



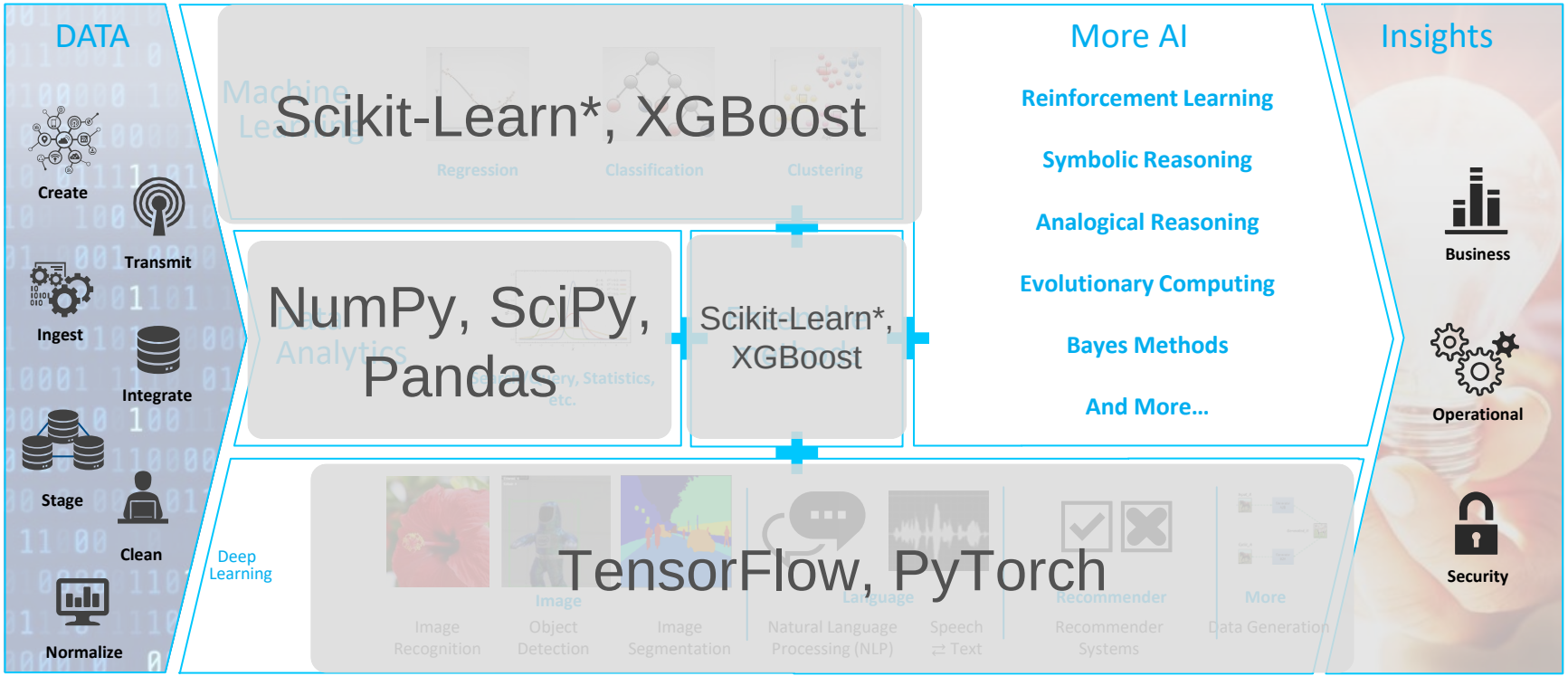
- Regression
- Classification
- Clustering
- Decision Trees
- Data Generation

- Image Processing
- Speech Processing
- Natural Language Processing
- Recommender Systems
- Adversarial Networks

AI is Interdisciplinary



AI is Interdisciplinary



AI is Everywhere…

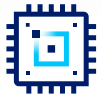
But not easy…

87% of AI models don’t move beyond POC

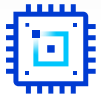
References:
https://venturebeat.com/2019/07/19/why-do-87-of-data-science-projects-never-make-it-into-production/
https://blogs.gartner.com/andrew_white/2019/01/03/our-top-data-and-analytics-predicts-for-2019/

AI Hardware Acceleration

GENERAL PURPOSE — PURPOSE BUILT



CPU



GPU

ACCELERATORS

intel CORE

intel XEON

Xe

intel MOVIDIUS

habana

Software AI Accelerators *(a.k.a. Intel's Magic AI Touch)*

HW Acceleration

Up to

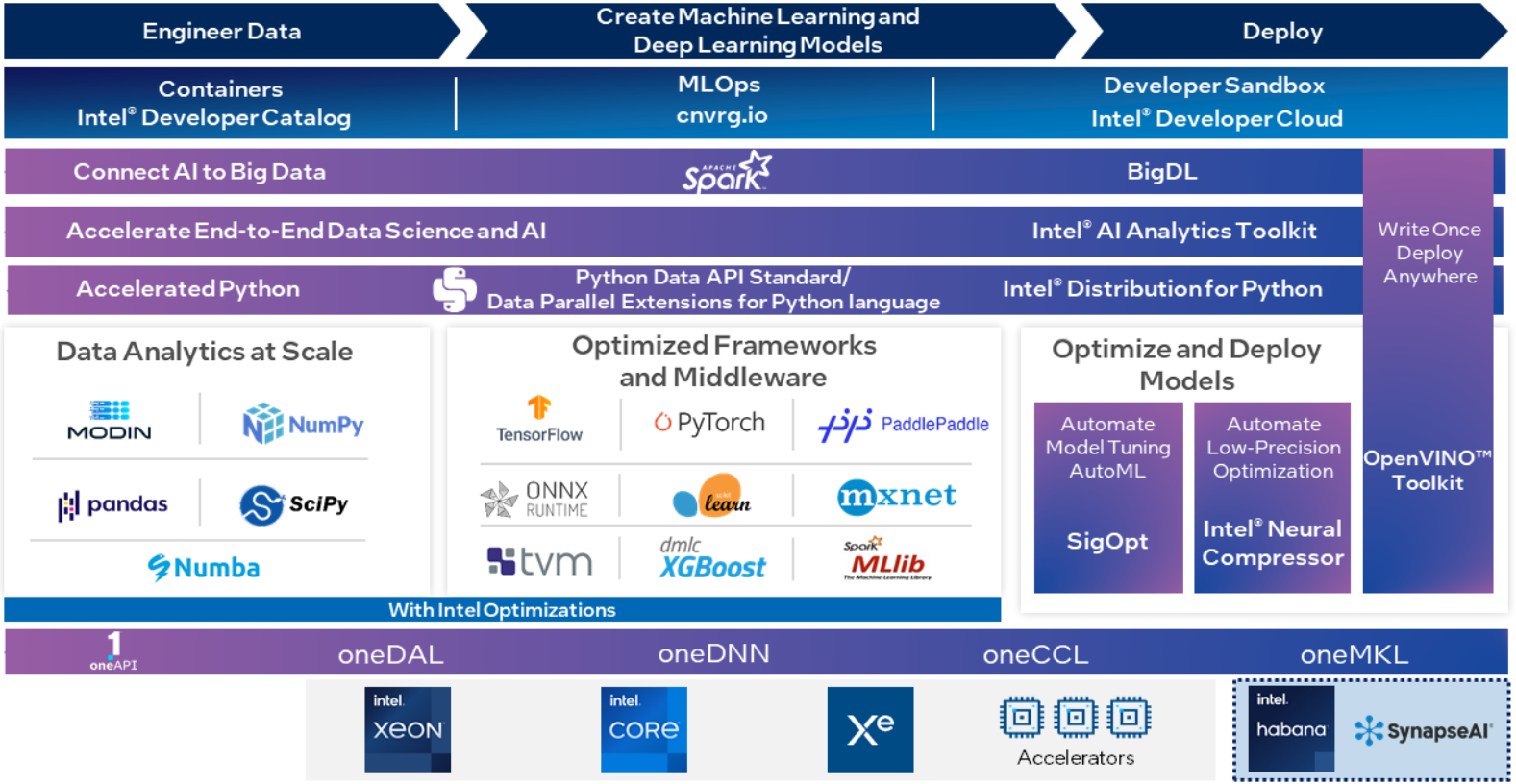
10 - 100x

With SW Acceleration

For details: https://venturebeat.com/2021/09/22/software-ai-accelerators-ai-performance-boost-for-free/

Source: Intel, https://venturebeat.com/2021/09/22/software-ai-accelerators-ai-performance-boost-for-free
Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex

AI Software Ecosystem and Intel Tools



* Other names and brands may be claimed as the property of others

Intel® Extension for Scikit-learn*

Common Scikit-learn*

```
from sklearn.svm import SVC

X, Y = get_dataset()


clf = SVC().fit(X, y)
res = clf.predict(X)
```

scikit-learn* mainline

Scikit-learn* with Intel CPU opts

```
from sklearnex import patch_sklearn
patch_sklearn()

from sklearn.svm import SVC

X, Y = get_dataset()


clf = SVC().fit(X, y)
res = clf.predict(X)
```

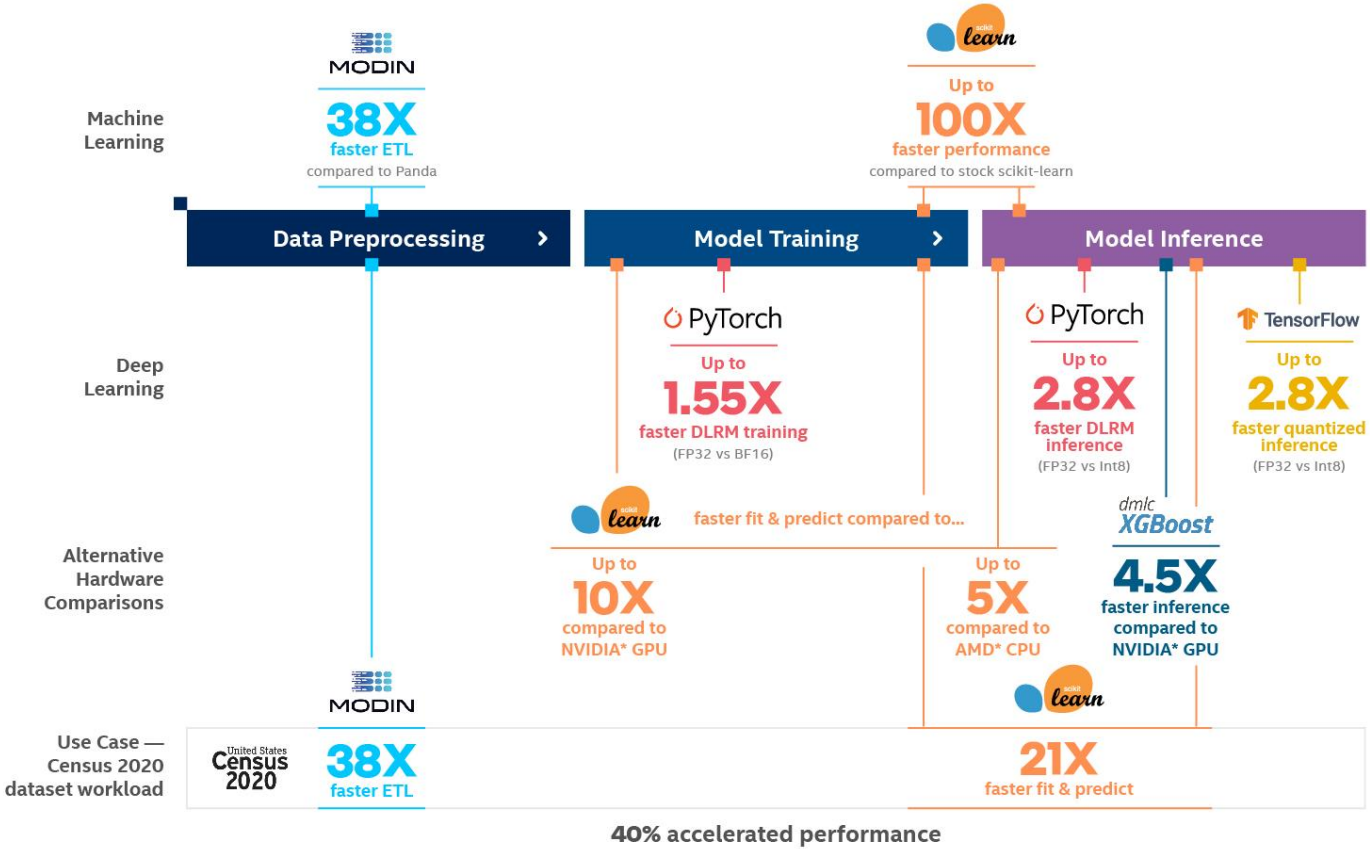
Same Code, Same Behavior

PASSED

- scikit-learn*, not scikit-learn*-like
- scikit-learn* conformance (mathematical equivalence) defined by scikit-learn* Consortium, continuously vetted by public CI

Available through:

- conda install scikit-learn-intelex
- conda install –c intel scikit-learn-intelex
- conda install –c conda-forge scikit-learn-intelex
- pip install scikit-learn-intelex

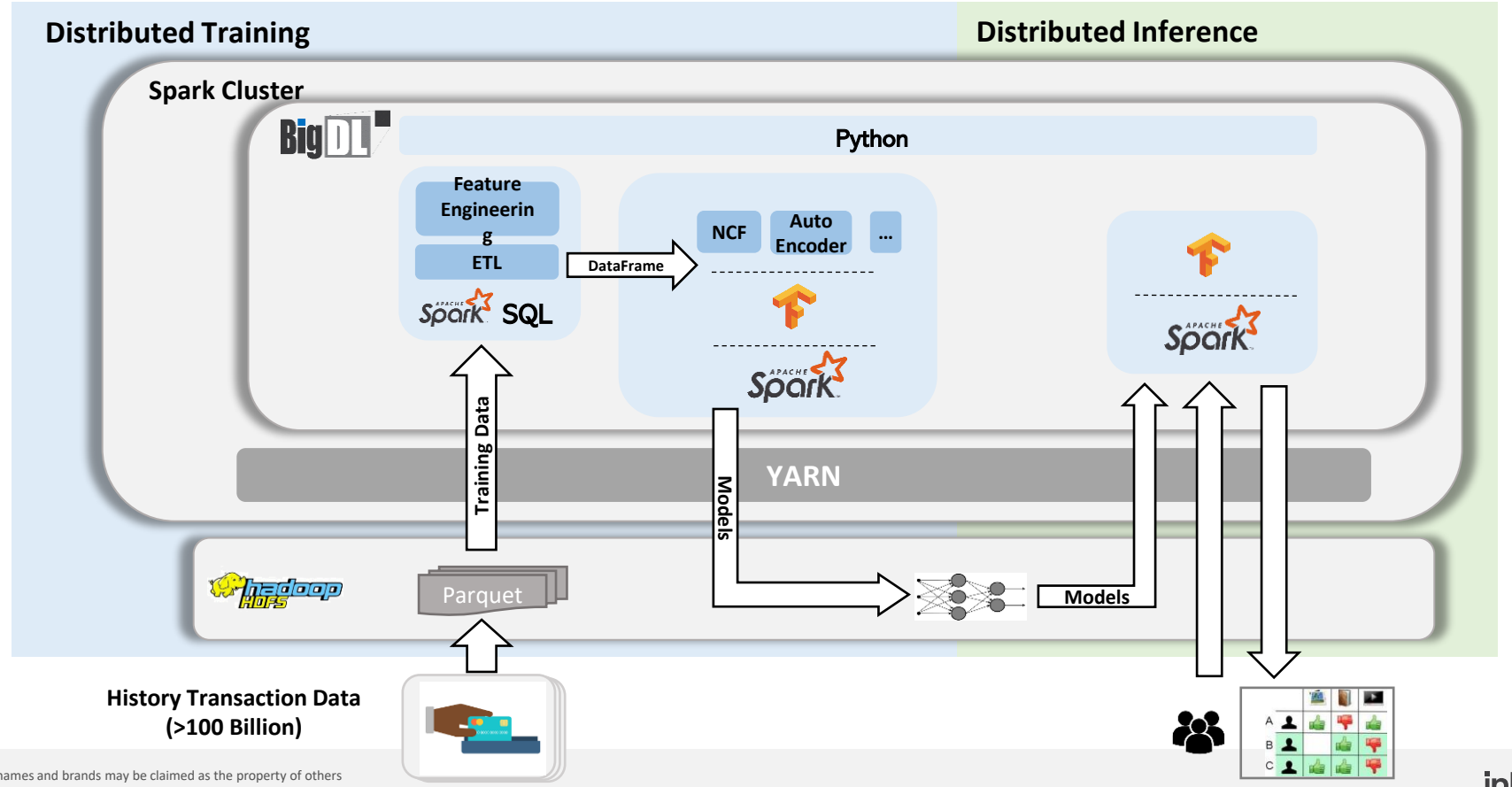


*Performance improvements shown here are based off hardware running on Intel Cascade Lake processors. This chart will be updated once data from Ice Lake is available. See backup for workloads and configurations. Results may vary.

** Not STAC benchmarks

Mastercard Recommender AI Service

End-to-End users-items propensity modelling pipeline on



AI at Scale with Mastercard and BigDL

www.intel.com/content/www/us/en/developer/articles/technical/ai-at-scale-in-mastercard-with-bigdl0.html

Customer Success

"Analytics Zoo, the Intel® AI Analytics Toolkit with the Intel® oneAPI Data Analytics library (oneDAL) helped **reduce end-to-end data processing time by 3x** and improved our **prediction model's accuracy by 2.5x** for AsiaInfo 5G network intelligence including customer satisfaction analysis, power saving for 5G base station and user location analysis." Duozhi Zhu, GM of 5G Network product R&D department, AsiaInfo Technologies Limited

"The Intel oneAPI Base and AI Analytics toolkits improved our 3D model reconstruction's **performance by up to 9x** on Intel® Xeon® platform compared to other GPU solutions." Daspatial, R&D General Manager Mr. Gao

"Allegro AI is a pioneer in machine learning and deep learning software platforms and tools. With Allegro AI customers deploy higher quality products, faster and more cost effectively. By integrating Intel's Intel oneAPI Data Analytics Library (oneDAL) and Intel AI Analytics Toolkit tools into Allegro Trains, Allegro AI offers a **better performance, and optimized use of cloud instances.**" Moses Guttman, Allegro AI CTO

"The Intel AI Analytics Toolkit's PyTorch 1.6 built using Intel oneAPI Deep Neural Network Library delivered **up to 11.4x[1] faster inferencing** for digital pathology medical screening." Director KFBIO

"With the help of Intel, we were able to **train, optimize, and deploy** a machine learning model **in less time and at a lower operational cost** than available alternatives, enabling us to get to market fast with a powerful solution that's optimized for Intel® architecture." Moloti Nakampe, R&D Director

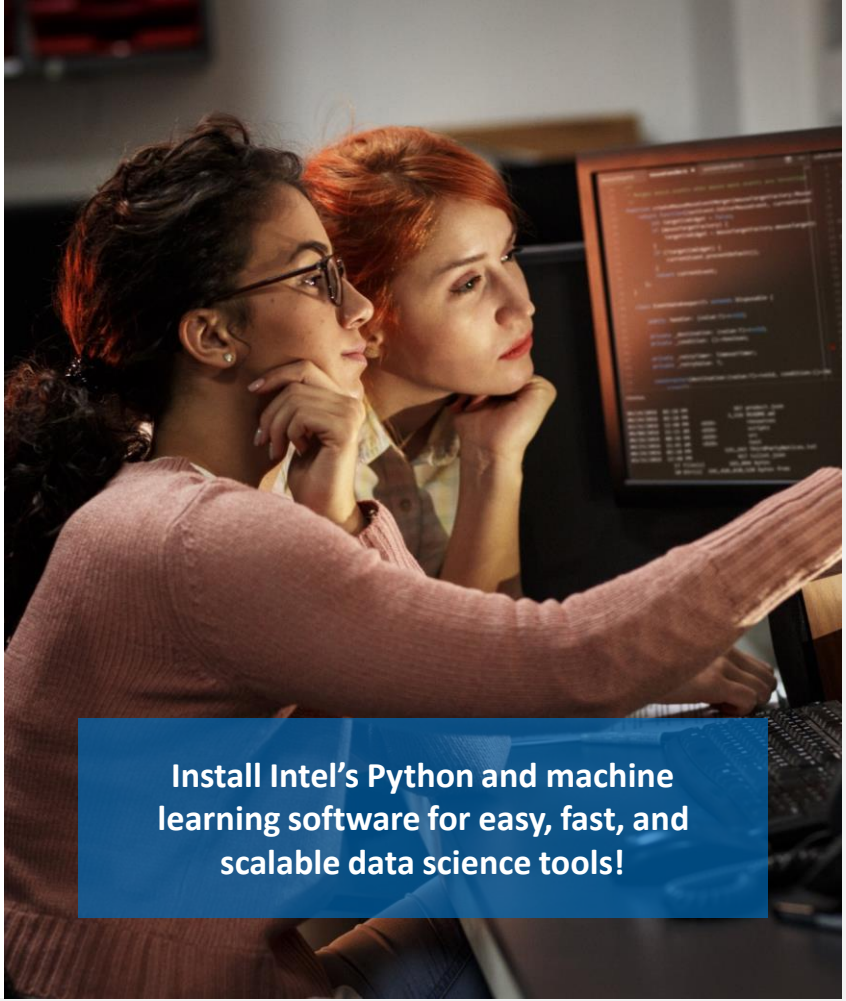
OEMs and system Integrators want to offer a competitive bundle to Nvidia RAPIDS for their AI workstations.
They are working with Intel to bundle Intel oneAPI AI Analytics Toolkit on Intel processor-based workstation systems.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy. Learn more at https://www.intel.com/content/www/us/en/developer/tools/oneapi/ecosystem-support.html

Call to Action

For more details on specific Intel's Python* AI and Data Science software options, visit

- Intel oneContainer Portal
- Intel® AWS Containers
- Intel® oneAPI AI Analytics Toolkit Code Samples
- Intel® Distribution for Python Support Forum
- Machine Learning and Data Analytics Support Forum
- Intel AI Homepage



Notices and Disclaimers

- Performance varies by use, configuration and other factors.
  Learn more at www.Intel.com/PerformanceIndex.
- Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.
- Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.
- Your costs and results may vary.
- Intel technologies may require enabled hardware, software or service activation.
-

Workloads and Configurations

See all benchmarks and configurations: https://software.intel.com/content/www/us/en/develop/articles/blazing-fast-python-data-science-ai-performance.html. Each performance claim and configuration data is available in the body of the article listed under sections 1, 2, 3, 4, and 5. Please also visit this page for more details on all scores, and measurements derived.

Testing Date: Performance results are based on testing by Intel as of October 16, 2020 and may not reflect all publicly available updates. **Configurations details and Workload Setup**: 2 x Intel® Xeon® Platinum 8280 @ 28 cores, OS: Ubuntu 19.10.5.3.0-64-generic Mitigated 384GB RAM (192 GB RAM (12x 32GB 2933). SW: Modin 0.81. Scikit-learn 0.22.2. Pandas 1.01, Python 3.8.5, DAL(DAAL4Py) 2020.2, Census Data, (21721922.45) Dataset is from IPUMS USA, University of Minnesota, www.ipums.org [Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Matthew Sobek. IPUMS USA: Version 10.0 [dataset], Minneapolis, MN. IPUMS, 2020. https//doc.org/10.18128/D010.V10.0]

Testing Date: Performance results are based on testing by Intel® as of October 23, 2020 and may not reflect all publicly available updates. **Configuration Details and Workload Setup**: Intel® oneAPI Data Analytics Library 2021.1 (oneDAL). Scikit-learn 0.23.1, Intel® Distribution for Python 3.8; Intel® Xeon® Platinum 8280LCPU @ 270GHz, 2 sockets, 28 cores per socket, 10M samples, 10 features, 100 clusters, 100 iterations, float32.

Testing Date: Performance results are based on testing by Intel® as of October 23, 2020 and may not reflect all publicly available updates. **Configuration Details and Workload Setup**: Intel® oneAPI AI Analytics Toolkit v2021.1; Intel® oneAPI Data Analytics Library (oneDAL) beta10, Scikit-learn 0.23.1, Intel® Distribution for Python 3.7, Intel® Xeon® Platinum 8280 CPU @ 2.70GHz, 2 sockets, 28 cores per socket, microcode: 0x4003003, total available memory 376 GB, 12X32GB modules, DDR4. **AMD Configuration**: AMD Rome 7742 @2.25 GHz, 2 sockets, 64 cores per socket, microcode: 0x8301038, total available memory 512 GB, 16X32GB modules, DDR4, oneDAL beta10, Scikit-learn 0.23.1, Intel® Distribution for Python 3.7. **NVIDIA Configuration**: NVIDIA Tesla V100 – 16 Gb, total available memory 376 GB, 12X32GB modules, DDR4, Intel® Xeon Platinum 8280 CPU @ 2.70GHz, 2 sockets, 28 cores per socket, microcode: 0x5003003, cuDF 0.15, cuML 0.15, CUDA 10.2.89, driver 440.33.01, Operation System: CentOS Linux 7 (Core), Linux 4.19.36 kernel.

Testing Date: Performance results are based on testing by Intel® as of October 13, 2020 and may not reflect all publicly available updates. **Configurations details and Workload Setup**: CPU: c5.18xlarge AWS Instance (2 x Intel® Xeon® Platinum 8124M @ 18 cores. OS: Ubuntu 20.04.2 LTS, 193 GB RAM. GPU: p3.2xlarge AWS Instance (GPU: NVIDIA Tesla V100 16GB, 8 vCPUs, OS: Ubuntu 18.04.2LTS, 61 GB RAM. SW: XGBoost 1.1: build from sources compiler – G++ 7.4, nvcc 9.1 Intel® DAAL: 2019.4 version: Python env: Python 3.6, Numpy 1.16.4, Pandas 0.25 Scikit-learn 0.21.2.

Workloads and Configurations

Testing Date: Performance results are based on testing by Intel® as of October 26, 2020 and may not reflect all publicly available updates. **Configuration Details and Workload Setup**: Intel® Optimization for Tensorflow v2.2.0; oneDNN v1.2.0; Intel® Low Precision Optimization Tool v1.0; Platform; Intel® Xeon® Platinum 8280 CPU; #Nodes 1; #Sockets: 2; Cores/socket: 28; Threads/socket: 56; HT: On; Turbo: On; BIOS version:SE5C620.86B.02.01.0010.010620200716; System DDR Mem Config: 12 slots/16GB/2933; OS: CentOS Linux 7.8; Kernel: 4.4.240-1.el7.elrepo x86_64.

Testing Date: Performance results are based on testing by Intel® as of February 3, 2021 and may not reflect all publicly available updates. **Configuration Details and Workload Setup:** Intel® Optimization for PyTorch v1.5.0; Intel® Extension for PyTorch (IPEX) 1.1.0; oneDNN version: v1.5; DLRM: Training batch size (FP32/BF16): 2K/instance, 1 instance; DLRM dataset (FP32/BF16): Criteo Terabyte Dataset; BERT-Large: Training batch size (FP32/BF16): 24/Instance. 1 Instance on a CPU socket. Dataset (FP32/BF16): WikiText-2 [https://www.salesforce.com/products/einstein/ai-research/the-wiktext-dependency-language-modeling-dataset/]: ResNext101-32x4d: Training batch size (FP32/BF16): 128/Instance, 1 instance on a CPU socket, Dataset (FP32/BF16): ILSVRC2012; DLRM: Inference batch size (INT8): 16/instance, 28 instances, dummy data. Intel® Xeon® Platinum 8380H Processor, 4 socket, 28 cores HT On Turbo ON Total memory 768 GB (24 slots/32GB/3200 MHz), BIOS; WLYDCRBLSYS.0015.P96.2005070242 (ucode: OX 700001b), Ubuntu 20.04 LTS, kernel 5.4.0-29-genen: ResNet50: [https://github.com/Intel/optimized-models/tree/master/pytorch/ResNet50]: ResNext101 32x4d: [https://github.com/intel/optimized-models/tree/master/pytorch/ResNext101_32x4ct: DLRM: https//github.com/intel/optimized-models/tree/master/pytorch/dlrm].

Testing Date: Performance results are based on testing by Intel® as of October 4, 2021 and may not reflect all publicly available updates. **Configuration Details and Workload Setup**: **Hardware (same for all configurations):** 1-node, 2x 2nd Gen Intel® Xeon® Gold 6258R on Lenovo 30BC003DUS with 768GB (12 slots/ 64GB/ 2666) total DDR4 memory and 2TB (4 slots/ 512GB/ 2666) DCPMM memory, microcode 0x5003102, HT on, Turbo on, Ubuntu 20.04.3 LTS, 5.10.0-1049-oem, 1x Samsung 1TB SSD OS Drive, 4x Samsung 2TB SSD in RAID0 data drive, 3x NVIDIA Quadro RTX 8000. **3 months of NYCTaxi Data on Stock Software Configuration:** Python 3.9.7, Pandas 1.3.3, Scikit-Learn 1.0, XGBoost 0.81, IPython 7.28.0, IPKernel 6.4.1. **Full 30 months of NYCTaxi Data on Nvidia RAPIDS Software Configuration:** Python 3.7.10, Pandas 1.2.5, XGBoost 1.4.2, cuDF 21.08.03, cudatoolkit 11.2.72, dask-cudf 21.08.03, dask-cuda 21.08.00, IPython 7.28.0, IPKernel 6.4.1. **Full 30 months of NYCTaxi Data on Intel Optimized Software Configuration:** Python 3.9.7, Pandas 1.3.3, Modin 0.11.0, OmniSci 5.7.0, Scikit-:earn 1.0, Intel® Extension for Scikit-Learn* 2021.3.0, XGBoost 1.4.2, IPython 7.28.0, IPKernel 6.4.1. NYCTaxi Dataset from New York City (nyc.gov): [https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page]