2020年9月29日

Intel AIのこれまでとこれから

インテル株式会社

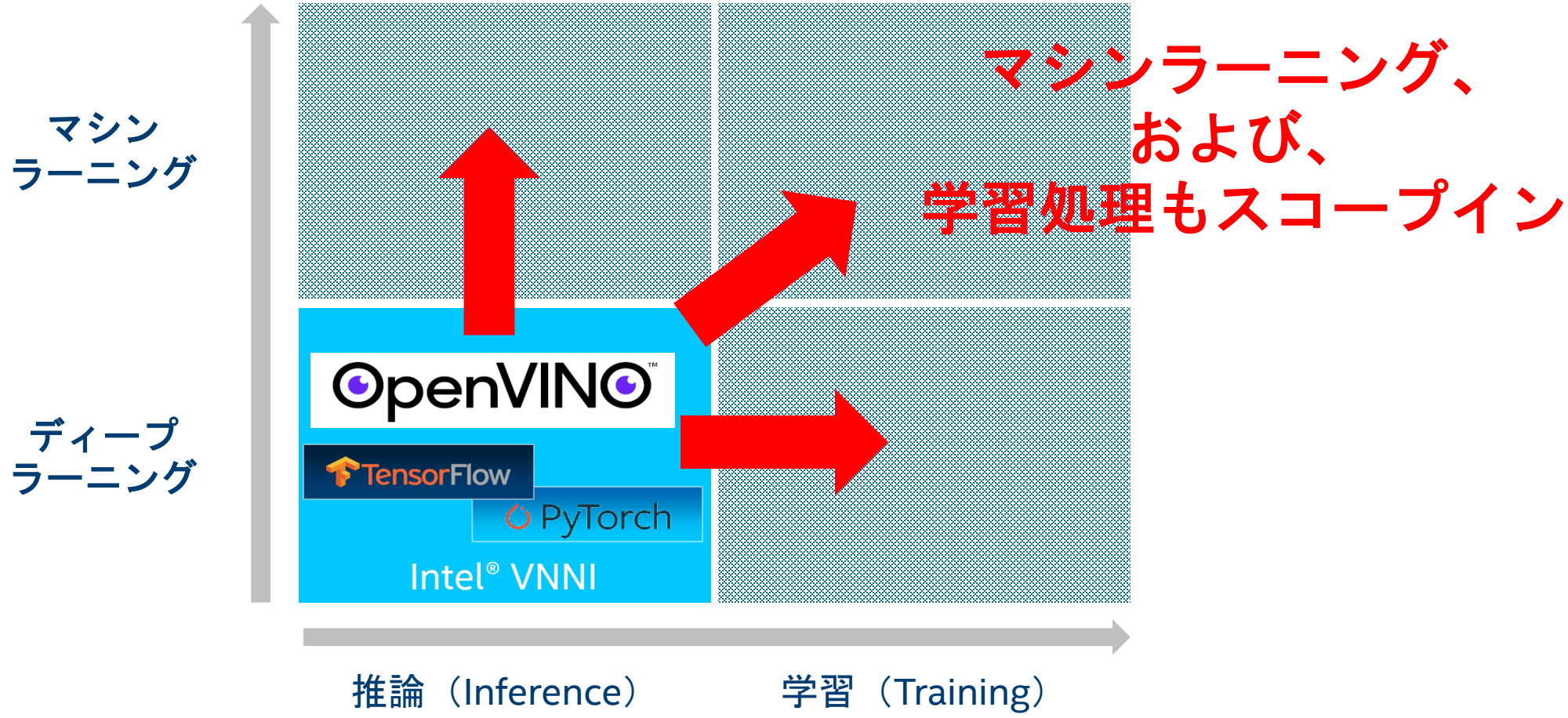
APJデータセンター・グループ・セールス

大内山　浩

AI "も" 動かすCPU

- あらゆるワークロードに対応できる汎用性と柔軟性がCPUの特徴

ここ最近のインテルAI
～より広いAIワークロードに対応～



マシンラーニング、
および、
学習処理もスコープイン

ディープラーニング高速化の要 ～AVX-512 & DL Boost～

- AVX-512（SIMD）が搭載され、並列演算性能の向上に寄与しております。更に、Deep Learning Boostという専用命令により更なるアクセラレーションが期待できます。

Intel® AVX-512

(Intel® Advanced Vector Extensions 512)

+

Intel® DL Boost

(Intel® Deep Learning Boost)

はいってる

第10世代Ice Lakeから搭載

はいってる

Skylake世代からAVX-512搭載
Cascade Lake世代からDL Boost搭載

インテル® AI ソフトウェア: マシンラーニングとディープラーニング

DEVELOPER TOOLS

App Developers

SW Platform Developer

TOPOLOGIES & MODELS

Data Scientist

FRAMEWORKS

Data Scientist

GRAPH

ML Performance Engineer

KERNEL

ML Performance Engineer

MACHINE LEARNING

ANALYTICS ZOO

- **Intel Distribution for Python** (SKlearn, Pandas)

python

- Intel Data Analytics Acceleration Library (Intel DAAL)
- Intel Math Kernel Library (Intel MKL)

DEEP LEARNING

OpenVINO

TensorFlow

PyTorch

- Intel Machine Learning Scaling Library (Intel MLSL)
- Intel® Deep Neural Network Library (DNNL)

MANAGEMENT TOOLS

Containers

kubernetes

Kubeflow

Deep Learning Reference Stack

Data Analytics Reference Stack

Architect & DevOps

CPU

CPU ■ GPU ■ FPGA ■ 専用

Red font products are the most broadly applicable SW products for AI users

インテルによるディープラーニング・フレームワークの最適化

スケール

- ロード・バランシング向上
- 同期イベント、all-to-all 通信の削減

全コアの有効活用

- OpenMP、MPI
- 同期イベント、直列コードの削減
- ロード・バランシング向上

ベクトル演算 / SIMD

- SIMD レーンごとのユニット・ストライド・アクセス
- 高いベクトル効率
- データ・アライメント

効率的なメモリーとキャッシュ利用

- ブロッキング
- データ再利用
- プリフェッチ
- メモリー・アロケーション

= TensorFlow, mxnet, Caffe2, PyTorch, BigDL for Spark

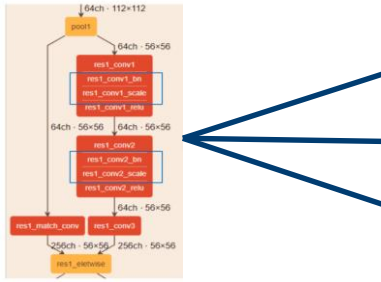
インストール・ガイドはこちら↓
ai.intel.com/framework-optimizations/

更なるフレームワークの最適化が進行中
(例、PaddlePaddle* 、CNTK* など)

SEE ALSO: Machine Learning Libraries for Python (Scikit-learn, Pandas, NumPy), R (Cart, randomForest, e1071), Distributed (MlLib on Spark, Mahout)
*Limited availability today
Optimization Notice

推論処理の更なる性能向上のための ディープラーニング・モデルの最適化と量子化

- 最適化：不要な Ops の除去、複数の Ops の統合などによりモデルをスマート化
- 量子化*：モデル内部の数値表現を FP32→INT8 に変換することでスリム化

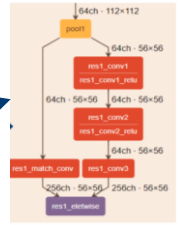


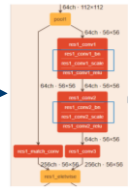
元のモデル
（TensorFlow*、PyTorch* などで作成）

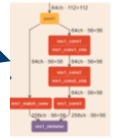
最適化

量子化

最適化 & 量子化







by OpenVINO™

by TensorFlow PyTorch

※各フレームワークごとに量子化ツールを用意

by OpenVINO™

* 2020年5月現在、インテル® ディープラーニング・ブースト (VNNI) が搭載された 第 2 世代インテル® Xeon® スケーラブル・プロセッサー以降、第 10 世代 インテル® Core™ プロセッサー・ファミリー (Ice Lake† のみ)以降 にてより効力を発揮する

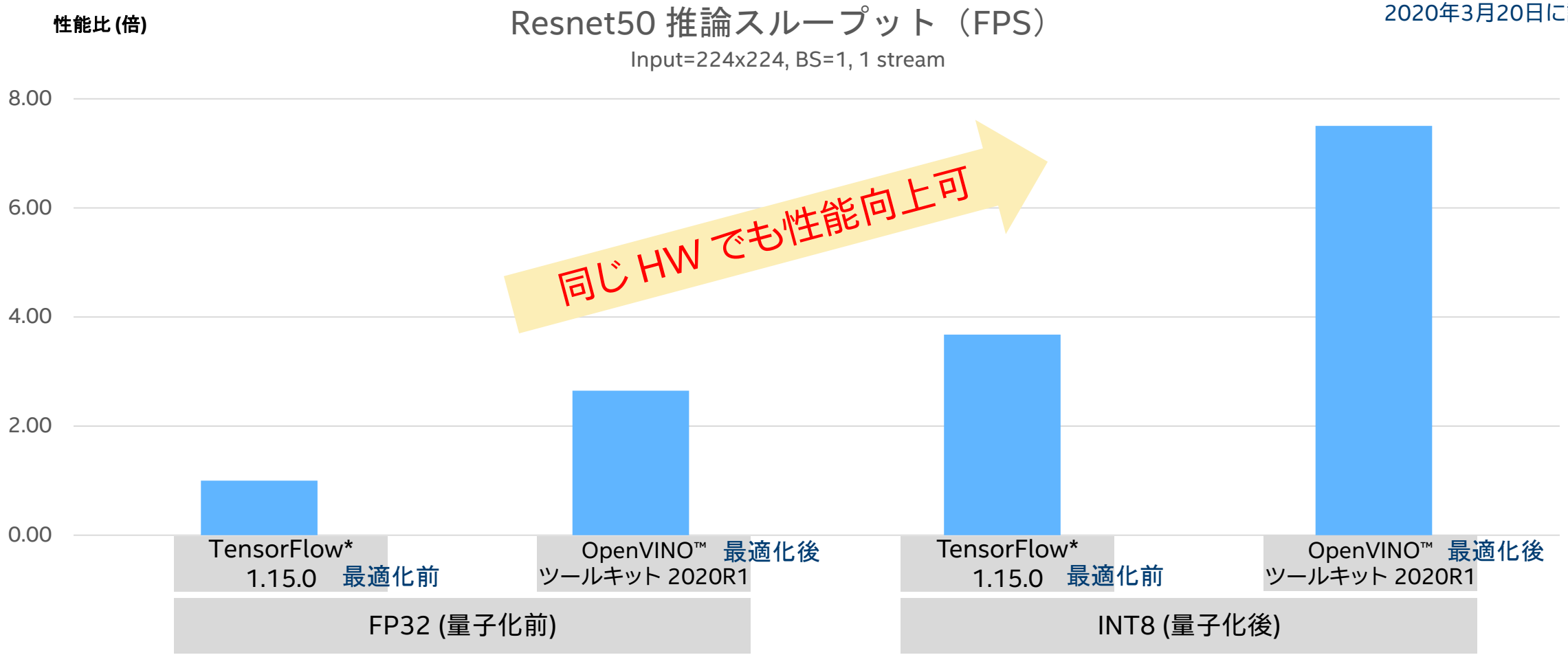
参考値

DL推論

ディープラーニング推論処理ベンチマーク

インテル® Xeon® Gold 6254 プロセッサー @ 2.10GHz (18 cores × 1 sockets)

2020年3月20日に計測



注)インテル社員による性能確認のための個人的なベンチマーク結果であり、インテルの公式結果ではありません。

DL推論

事例：理化学研究所 様 CheXNet の推論性能改善

Before Optimization←　→ After Optimization

約2.2万枚のテスト画像をバッチ処理で推論

11,177 sec (Baseline)

on
Xeon 6252 x2

※オリジナルモデルは PyTorch 1.2.0にて実装

x10.0

最適化

- モデルをONNXに変換
- OpenVINOのモデルオプティマイザーでONNX→IRへ変換
- OpenVINOの推論エンジン上で同期実行

1,116 sec

x3.1

量子化

- OpenVINOの量子化ツールにてIR内一部のレイヤーの数値表現をINT8へ変換（ツールのカスタマイズ含む）
- OpenVINOの推論エンジン上で同期実行（VNNI利用）

359 sec

x1.4

並列化

- OpenVINOの推論エンジン上で非同期実行（8並列で推論処理を実行）

251 sec

x 44.5
against Baseline

上記対応内容は下記Githubを参照
https://github.com/taneishi/CheXNet
（計算科学研究機構 種石 様のレポジトリ）

Training with Huge Memory ~U-Net Training by NUS~

DL学習

NUS National University of Singapore | Saw Swee Hock School of Public Hea

GPU-based env

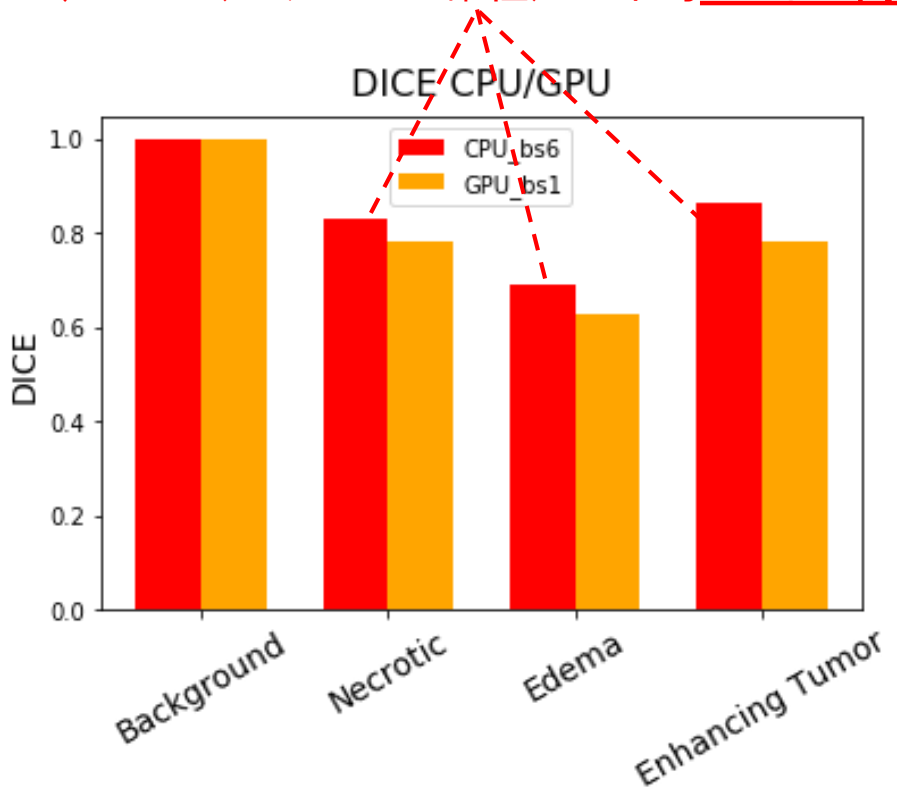
- V100 GPU (32GB memory)
- 10 CPU cores
- 126GB RAM
- **Batch size of 1**

CPU-based env

- 2 x Intel Platinum CPUs.
- 2 x 24 CPU cores
- 384GB RAM
- **Batch size of 6**

Result

インテル® CPU上でトレーニングしたモデルの方が、DICE（モデルの正確性）が平均**5%ほど高い**。



パフォーマンスが欲しい場合はどうすればいい?

↓

複数の CPU を束ねて使いましょう

つまり、分散学習 (Distributed Training) です

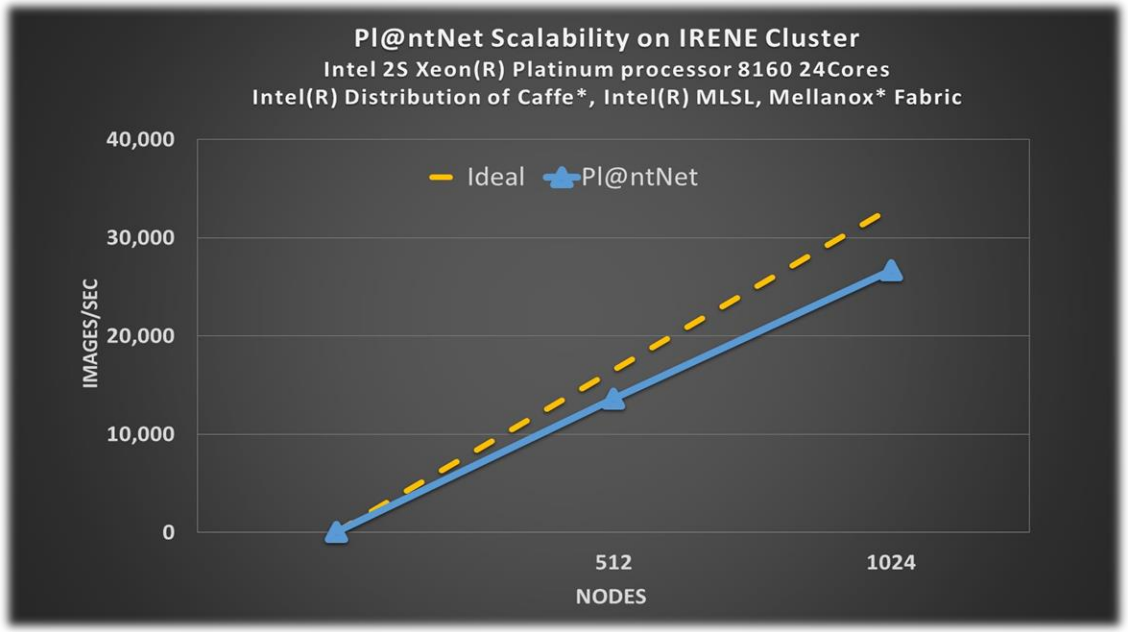
DL学習

既存インフラ上で効率的な深層学習のスケール ~GENCI と CERN の事例~

GENCI

French research institute focused on numerical simulation and HPC across all scientific and industrial fields

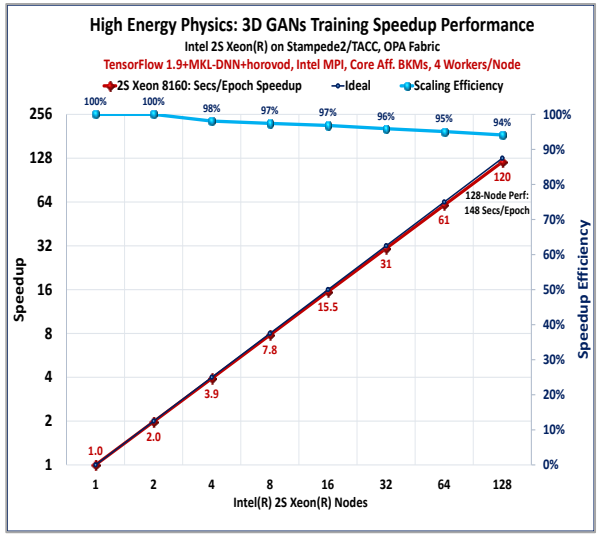
Succeeded in training a plant classification model for 300K species, 1.5TByte dataset of 12 million images on 1024 2S Intel® Xeon® Nodes with Resnet50.

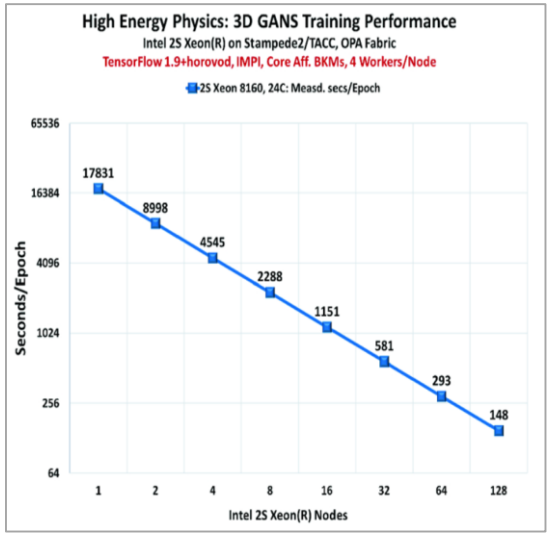


CERN

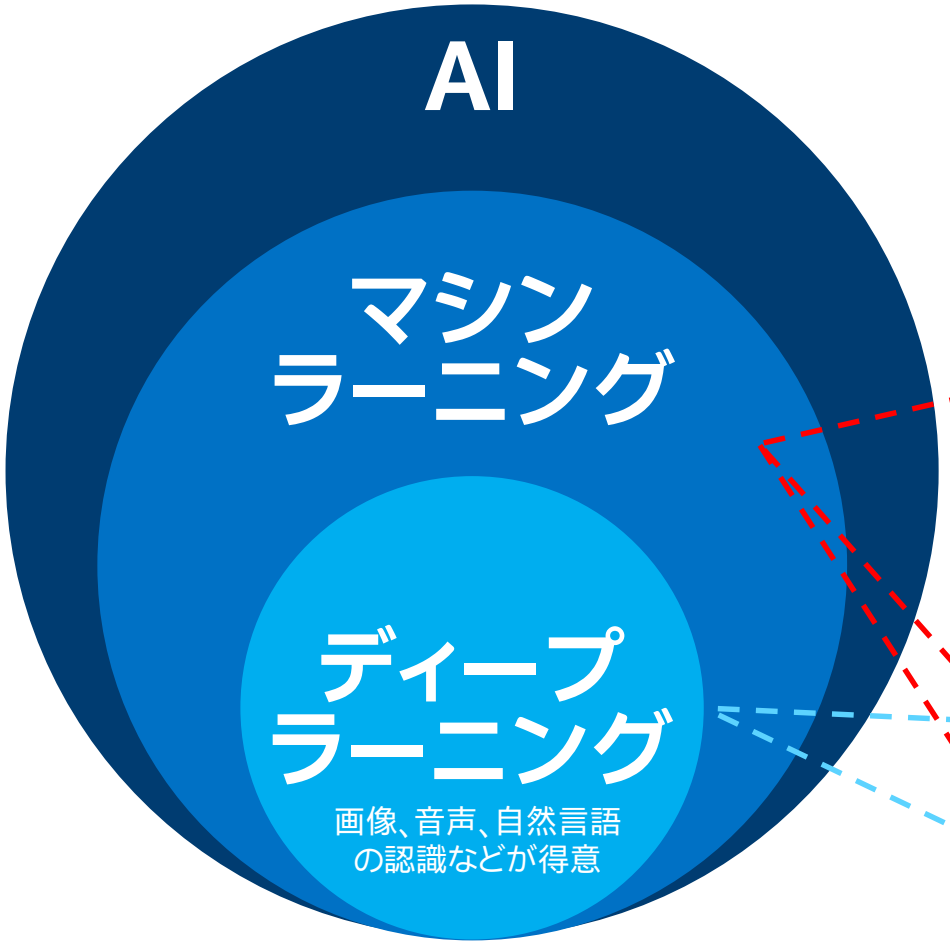
the European Organization for Nuclear Research, which operates the Large Hadron Collider (LHC), the world's largest particle accelerator

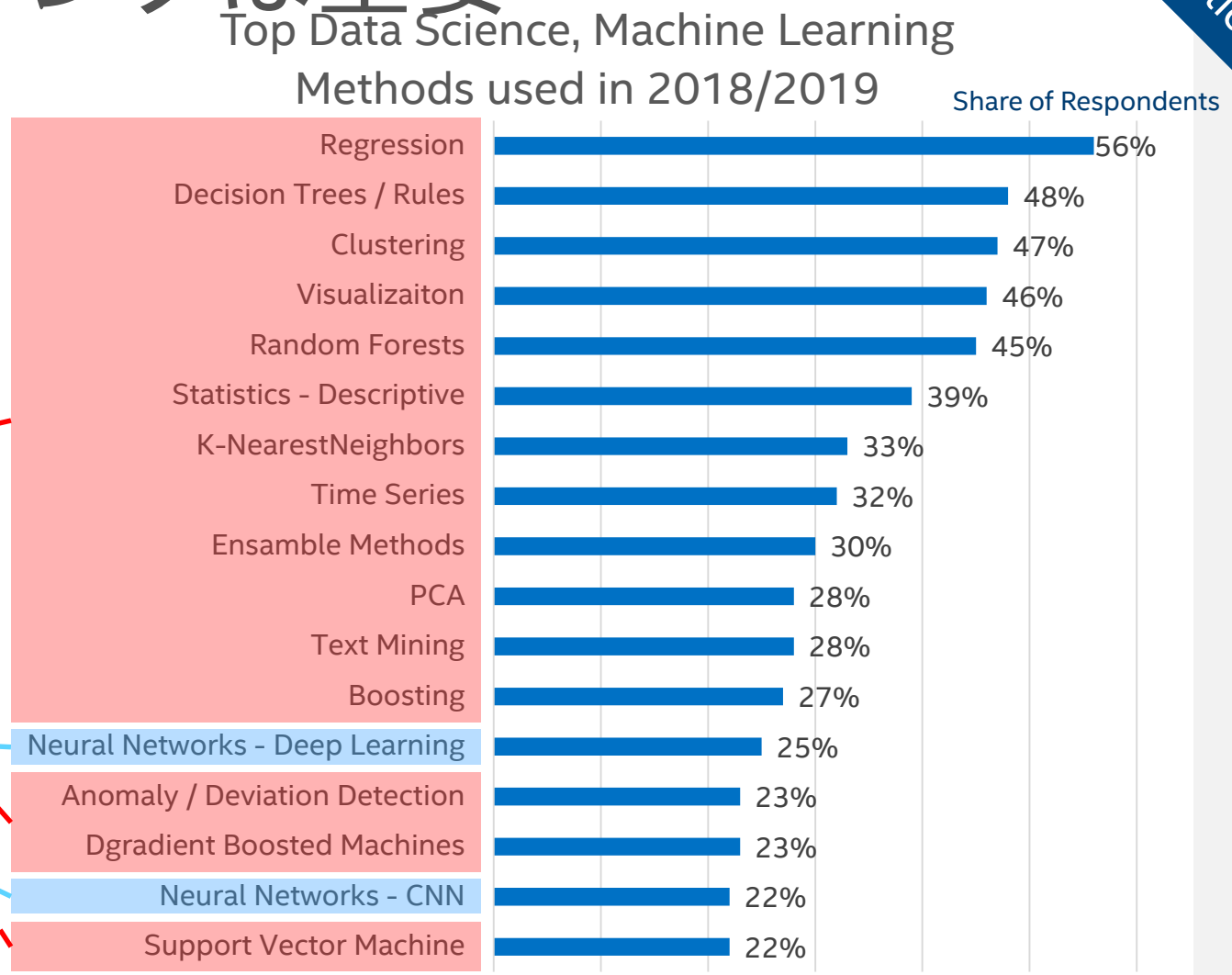
94% scaling efficiency up to 128 nodes, with a significant reduction in training time per epoch for 3D-GANs





まだまだマシンラーニングは重要





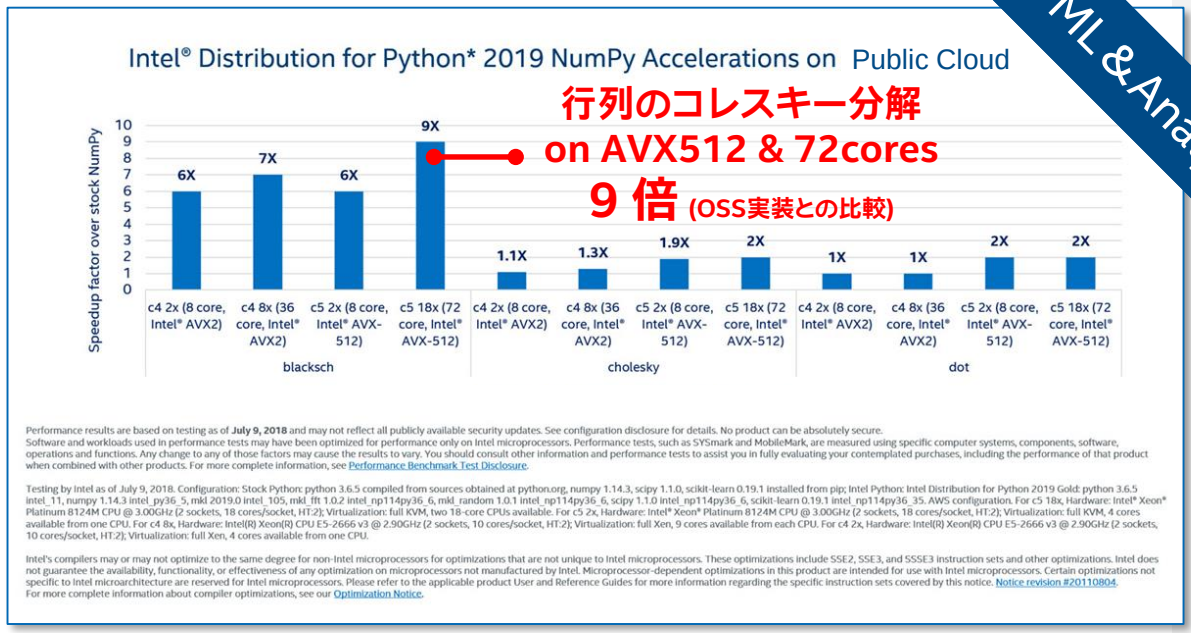
引用元：https://www.kdnuggets.com/2019/04/top-data-science-machine-learning-methods-2018-2019.html

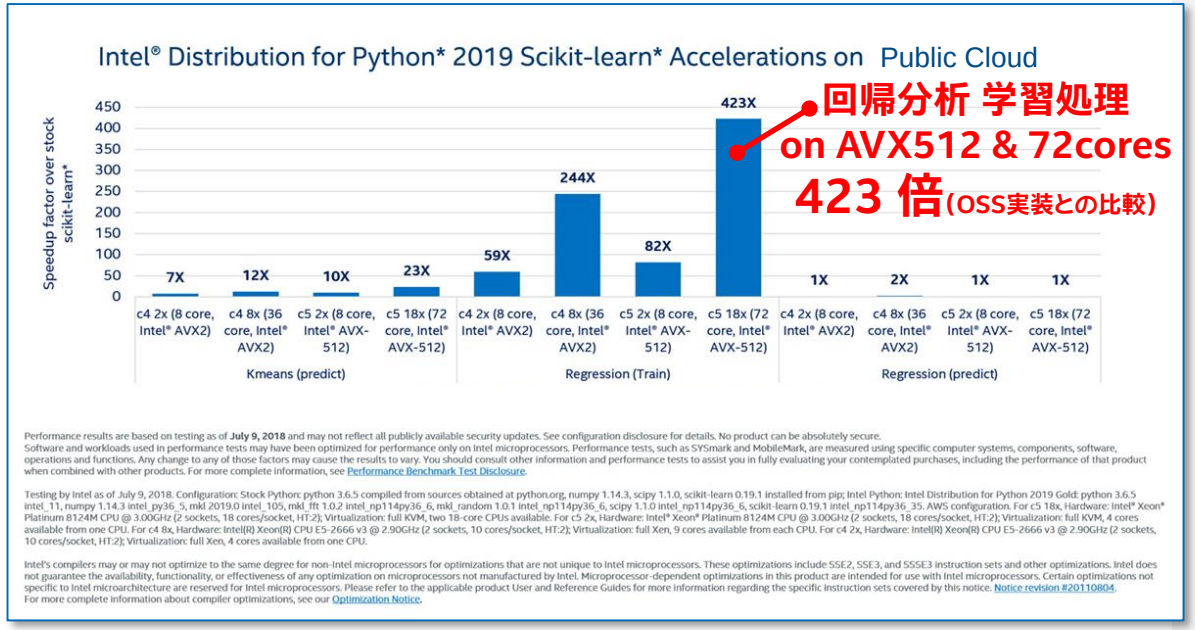
Intel® Distribution for Python*

インテルが実装、かつ、最適化した Python、および、周辺ライブラリ

- Numpy
- Pandas
- Scipy
- Scikit-learn
- XGBoost
- TensorFlow
- etc..

ML & Analytics





https://software.intel.com/en-us/distribution-for-python/benchmarks

インテル® のAI系ライブラリー & oneDALの使い方

ML & Analytics

数学	マシンラーニング／データ分析	ディープラーニング	集団通信
インテル® oneAPI Math Kernel Library (oneMKL)	インテル® oneAPI Data Analytics Library (oneDAL)	インテル® oneAPI Deep Neural Network Library (oneDNN)	インテル® oneAPI Collective Communication Library (oneCCL)

daal4py

scikit learn

APACHE Spark™

databricks

パートナー ソリューション

pip install daal4py	*pip install intel-scikit-learn*	*Sparkへのインストール可能*	*http://www.intel.com/analytics*

https://www.oneapi.com/

先進技術

新たなデマンドと新たなテクノロジー

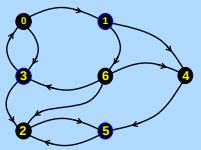
Security

PPML
(**P**rivacy **P**reserving **M**achine **L**earning)

プライバシー情報保護に重きを置いた機械学習技術

Data

Graph



グラフデータに対する分析または機械学習を用いてのパターン検出など

Algorithm

SLIDE
(**S**ub-**L**Inear **D**eep **l**earning **E**ngine)

ライス大学との共同研究。ディープラーニング学習アルゴリズムを抜本的に見直すことでCPUにてGPUを上回る学習性能を実現

グラフ分析に関するインテルの技術ブログ

先進技術

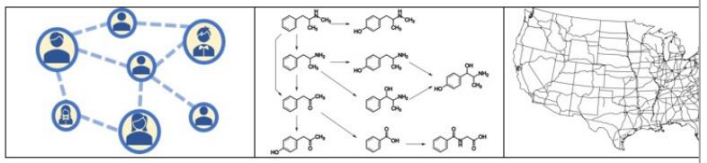
Measuring Graph Analytics Performance

The Diverse Landscape of Graph Analytics Requires a Comprehens
Benchmark

Henry Gabb Follow
Feb 8 · 6 min read

What Is Graph Analytics And Why Does It Matter?

A graph is a good way to represent a set of objects and the relations be
them (**Figure 1**). Graph analytics is the set of techniques to extract
information from connections between entities.



Adventures in Graph Analytics Benchmarking

It's Important to Use a Benchmark for Its Intended Purpose

Henry Gabb Follow
Feb 18 · 5 min read

With all the attention graph analytics is getting lately, it's increasingly important to measure its performance in a comprehensive, objective, and reproducible way. I covered this in a previous article, in which I recommended using an off-the-shelf benchmark like the GAP Benchmark Suite from the University of California, Berkeley. There are other graph benchmarks, of course, like LDBC Graphalytics, but they can't beat GAP f ease of use. There's significant overlap between GAP and Graphalytics, bu the latter is an industrial-strength benchmark that requires a special software configuration.

Measuring Graph Analytics Performance

You Don't Have to Spend $800,000 to Compute PageRank

There's a Better Way to Do Large-Scale Graph Analytics

Henry Gabb Follow
Jul 8 · 3 min read

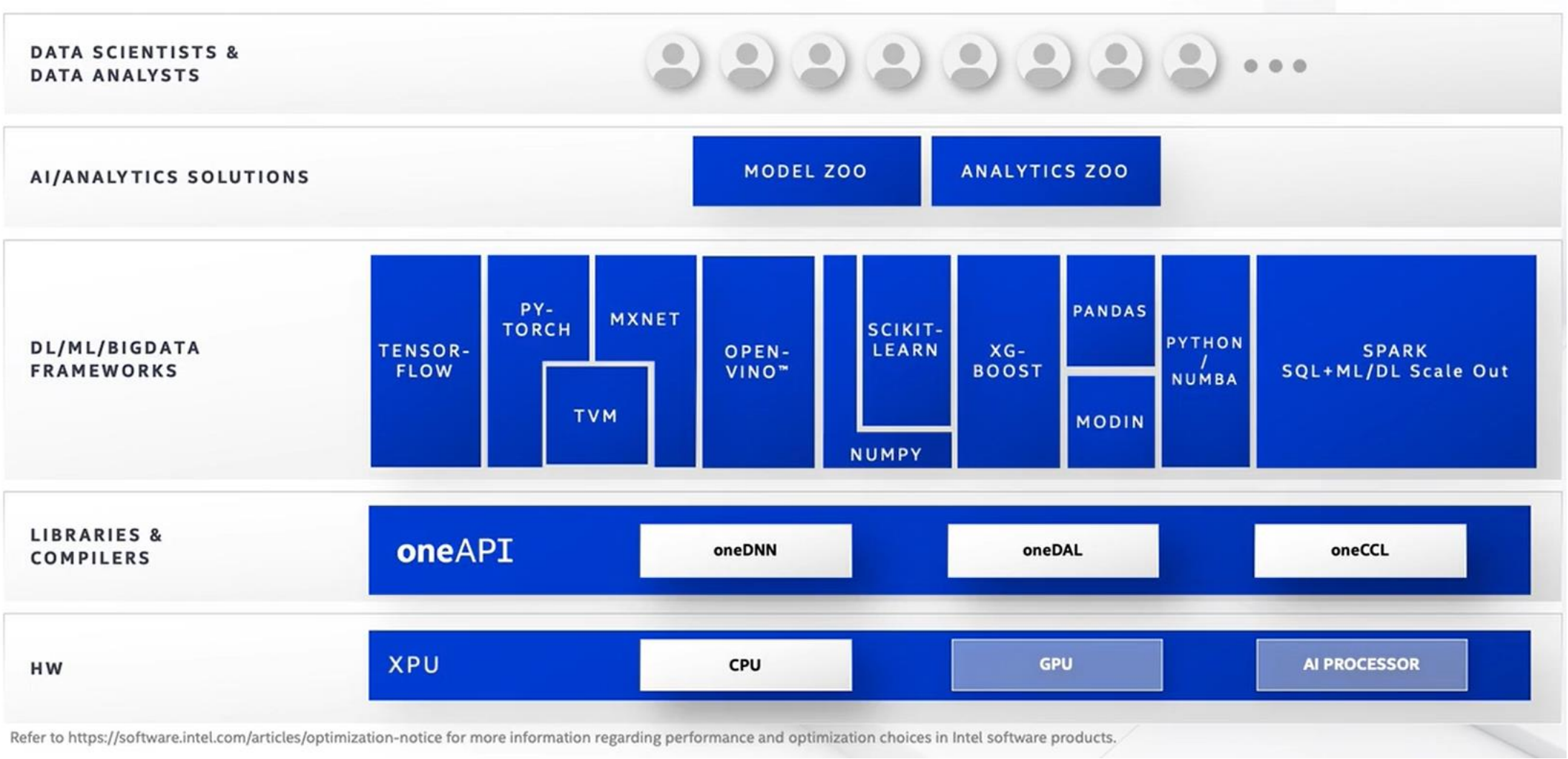
Benchmarking isn't my favorite topic, but I have a passing interest in graph analytics benchmarking:

Measuring Graph Analytics Performance
What Is Graph Analytics And Why Does It Matter?
medium.com

I'll occasionally dissect benchmarks that I think are inaccurate or misleading:

https://medium.com/intel-analytics-software

AI Software Ecosystem on Intel



Refer to https://software.intel.com/articles/optimization-notice for more information regarding performance and optimization choices in Intel software products.

インテル製品で AI への取り組みを加速

発見
可能性と次のステップ

データ
セットアップ、取り込み、クリーニング

開発
分析 / AI を使用するモデル

導入
実稼動へ & 反復

エコシステム

インテル® AI ビルダーズ・プログラム

100 以上の垂直的 / 水平的エコシステムのソリューション

最適化されたクラウド

Amazon Web Services*
Baidu* Cloud
Google Cloud* Platform
Microsoft* Azure*
など

AI に最適化された構成

intel select solution

ソフトウェア

データ分析

50 以上の最適化されたソフトウェア・プラットフォーム

マシンラーニング

Python* 向けインテル® ディストリビューション

ANALYTICS ZOO

ディープラーニング

TensorFlow　PyTorch

OpenVINO

ハードウェア

転送

intel Ethernet　BAREFOOT NETWORKS an Intel company

intel Silicon Photonics

格納

intel OPTANE DC PERSISTENT MEMORY　intel OPTANE DC SOLID STATE DRIVE

インテル® 3D NAND SSD

処理

intel CORE i7 10TH GEN　intel XEON　Xe　intel MOVIDIUS　habana An Intel Company

すべての製品、コンピューター・システム、日付、および数値は、現在の予想に基づくものであり、予告なく変更されることがあります。最適化に関する注意事項


intel