



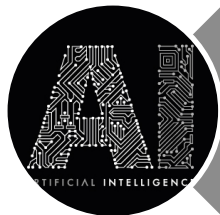
ARTIFICIAL INTELLIGENCE FOR CAPITAL MARKETS

Vin Sharma

vin.sharma@intel.com // @ciphr

10-18-2016

AGENDA



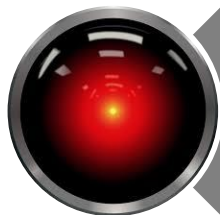
AI Era is Here... Again

- Why Now
- What's New



AI at Work

- Use Cases
- Challenges



AI Systems

- Hardware
- Software

WHY NOW?

Unstructured Data

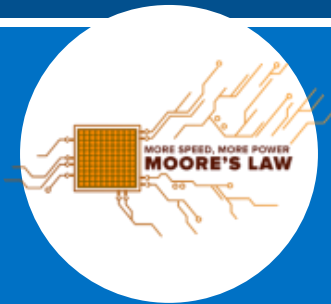


Image: 1000 KB / picture

Audio: 5000 KB / song

Video: 5,000,000 KB / movie

Better Hardware



Compute:

* Transistor density 2x /18 months

Storage:

* Cost / GB in 1995: \$1000.00

* Cost / GB in 2015: \$0.03

Smarter Algorithms



Advances in algorithms, including neural networks, leading to better accuracy in models that handle unstructured data

Artificial Intelligence

Sense, learn, reason, act, and adapt to the external world without explicit programming

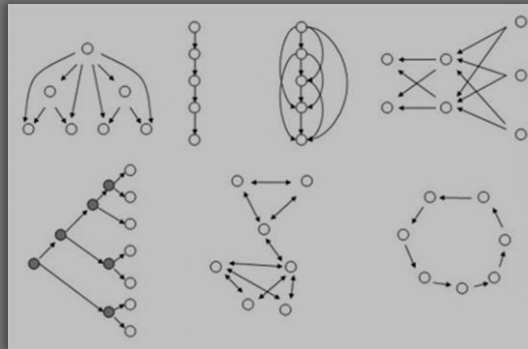
Perception

Detect or recognize patterns in audio, visual, tactile, ambient data



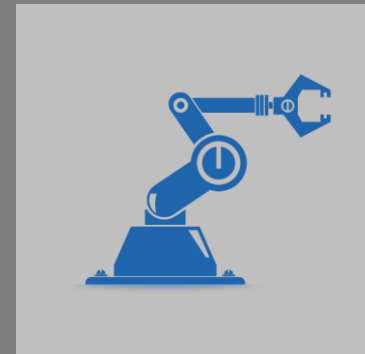
Cognition

Organize data patterns into meaningful structures and recommend action



Action

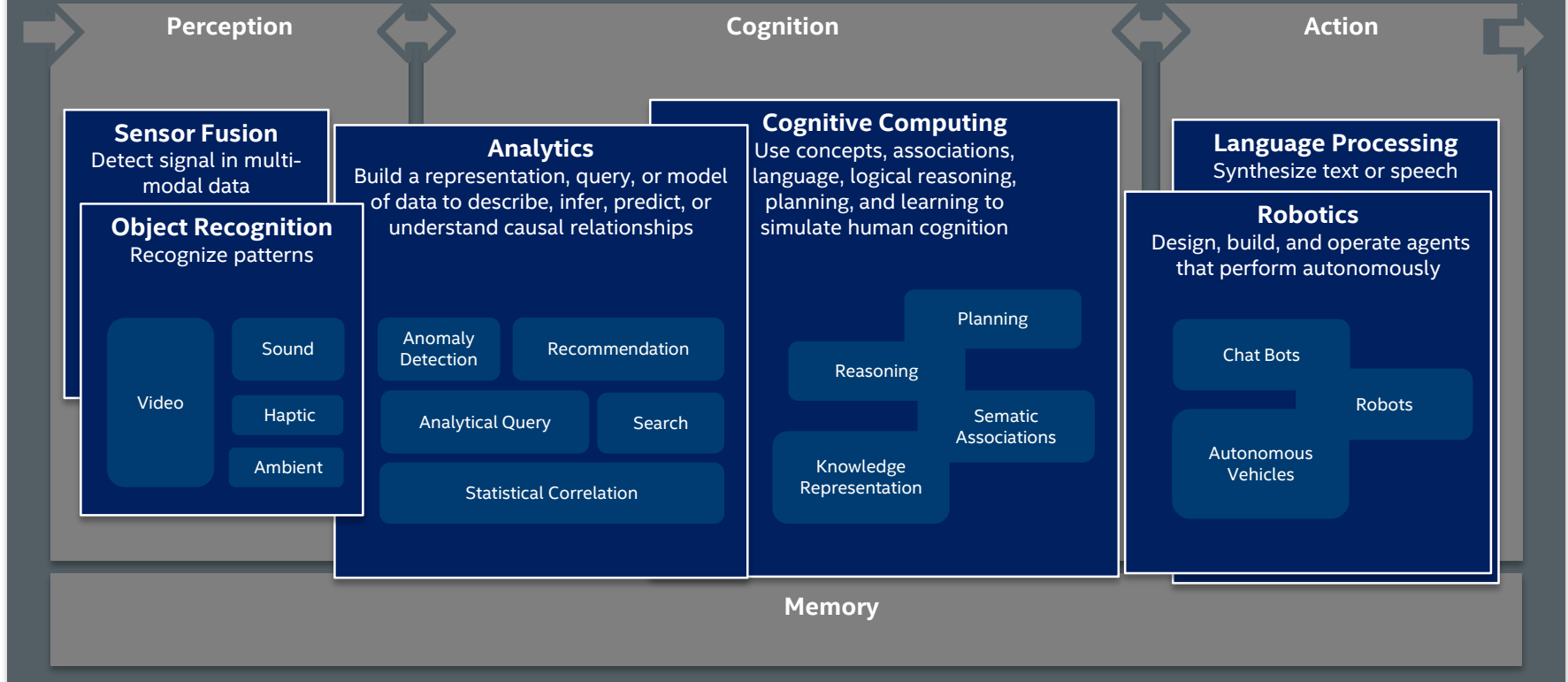
Communicate, control, or respond to external stimuli



Memory

Store and associate data, patterns, decisions, and actions for recall

Artificial Intelligence



Artificial Intelligence

Sense, perceive, reason, act, and adapt to the external world without explicit programming

Sensation → Perception

Detect or recognize patterns in audio, visual, tactile, ambient data

Pattern Analysis

Cognition

Organize data patterns into meaningful structures and recommend action

Knowledge Representation

Action

Communicate, control, or respond to external stimuli

Language

Robotics

Machine Learning

Computational methods that infer (with supervised, unsupervised, & reinforcement training) a predictive or causal model from data

Connectionist

Composition of nonlinear functions that learn successively complex representations

Statistical

Bayesian and other methods to improve statistical inference

Analogic

Use measures of similarity or distance to increment knowledge

Evolutionary

Use genetic methods to optimize fitness of a population or programs

Symbolic

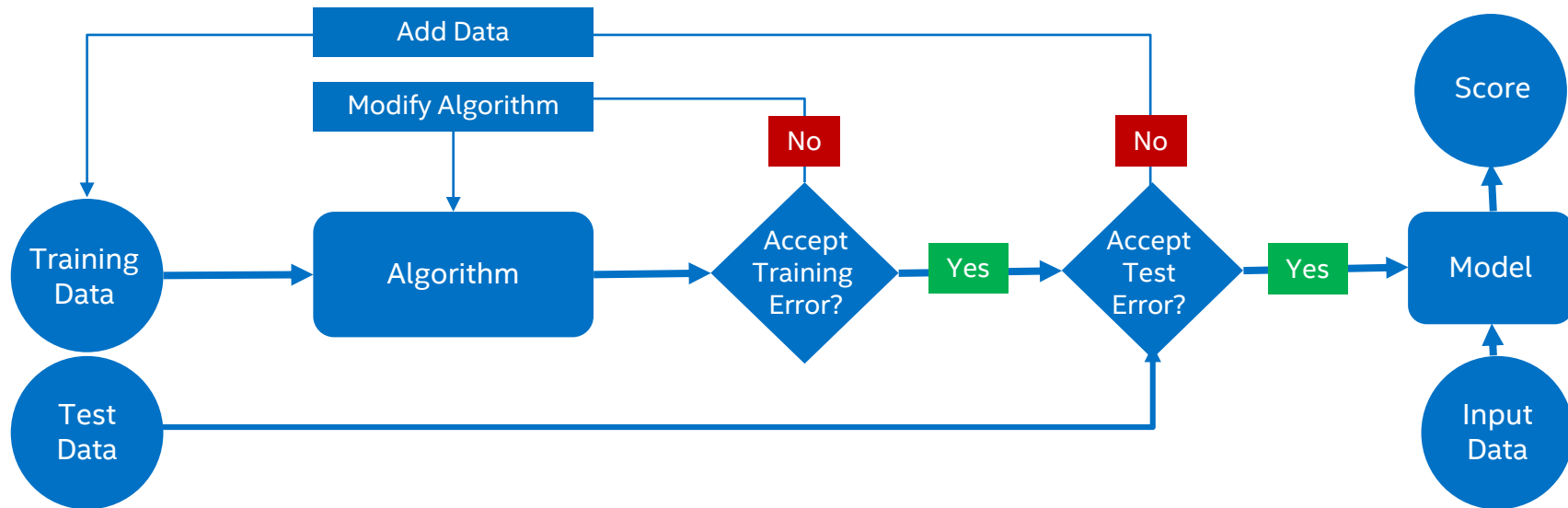
Use logic operations on symbols to deduce functions

Memory

Store and associate data, patterns, decisions, and actions for recall

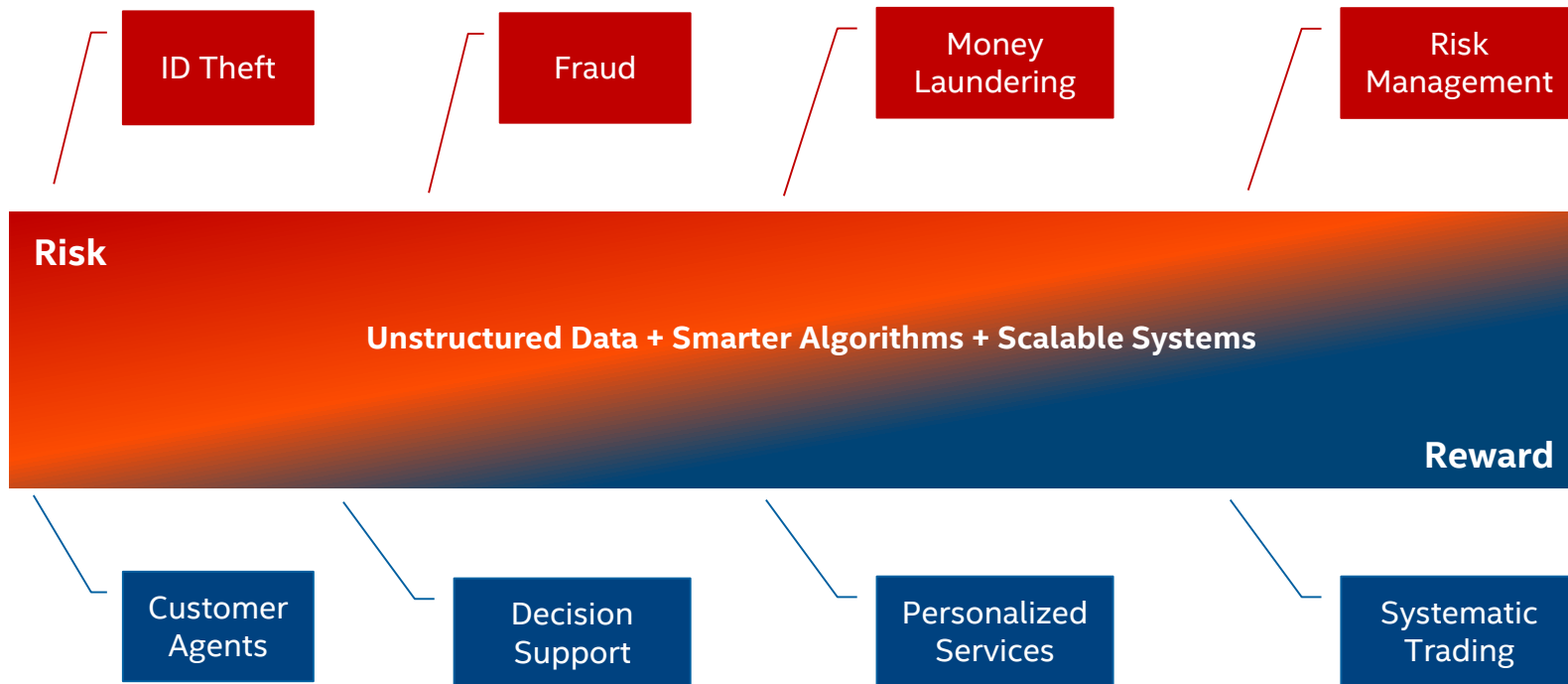
MACHINE LEARNING

Machine Learning is the study, development, and application of algorithms that improve their performance of some task based on experience (previous iterations).

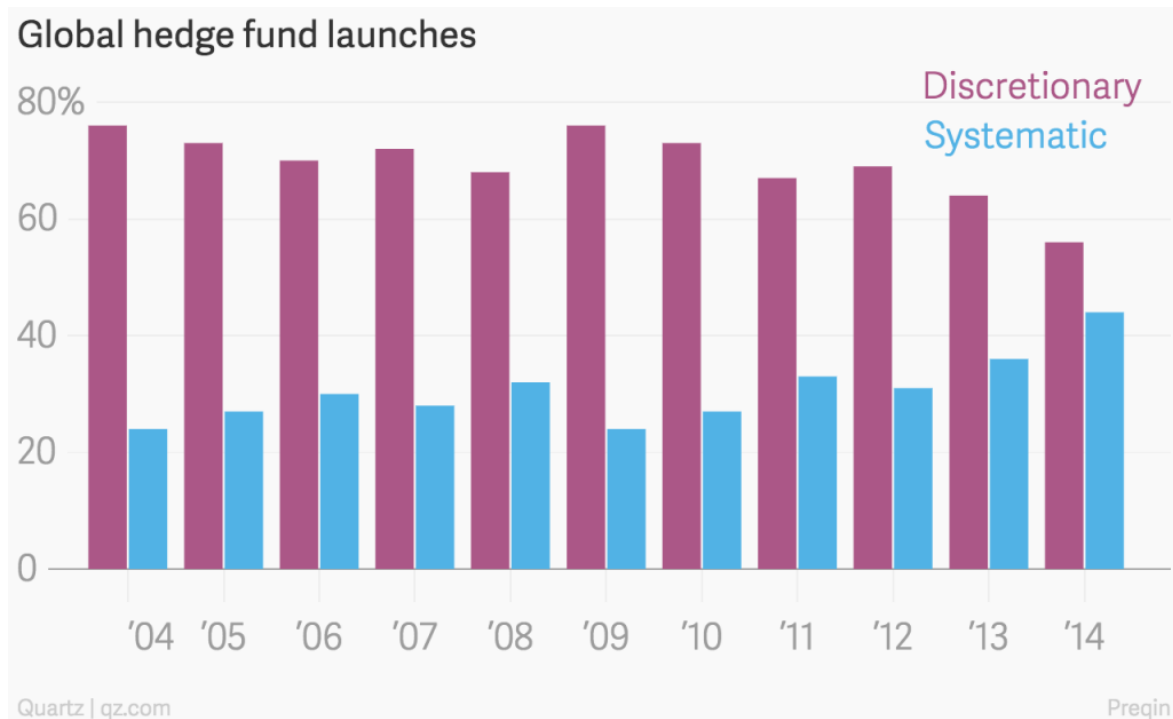


AI at Work

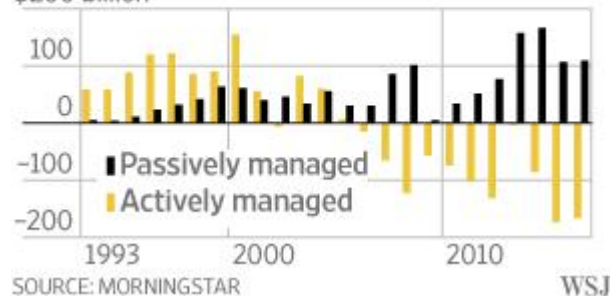
USE CASES: AI IN FSI



RISE OF SYSTEMATIC TRADING



Net flows of U.S. stock mutual and exchange-traded funds
\$200 billion



Artificial Intelligence

Sense, learn, reason, act, and adapt to the external world without explicit programming

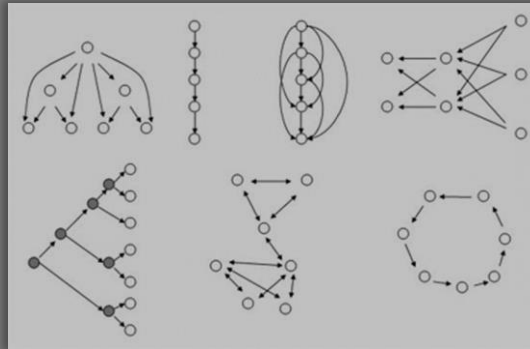
Perception

Detect or recognize patterns in audio, visual, tactile, ambient data



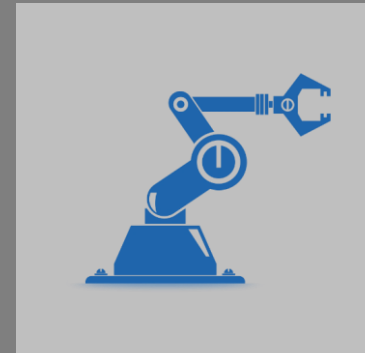
Cognition

Organize data patterns into meaningful structures and recommend action



Action

Communicate, control, or respond to external stimuli



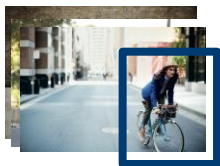
Memory

Store and associate data, patterns, decisions, and actions for recall

DEEP LEARNING

Step 1: Training (Over Hours/Days/Weeks)

Input data



Person

Create Deep network



Trained Model

Output Classification

90% person
8% traffic light

Step 2: Inference (Real Time)



New input from camera and sensors

Trained neural network model



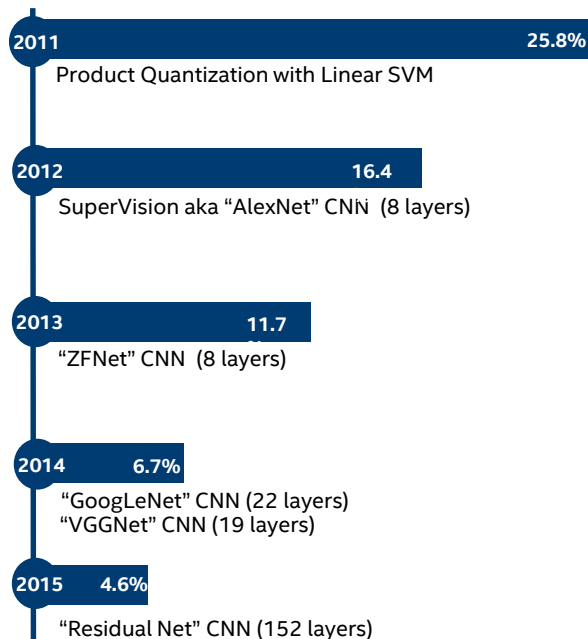
97% person

Output Classification

DEEP LEARNING FOR IMAGE CLASSIFICATION



CLASSIFICATION Top-5 Error Rate



Human-Level Accuracy

- Explicitly assumes that inputs are images
- Stack multiple convolution layers
- Other layers enable abstractions and classification
- Each conv layer consists of small spatial filters
- Filters learn successively complex representations
- Reuse reduces number of parameters dramatically
- Residual learning allows hundreds of layers
- Batch normalization maintains higher learning rate

CLASSIFICATION

Label the image

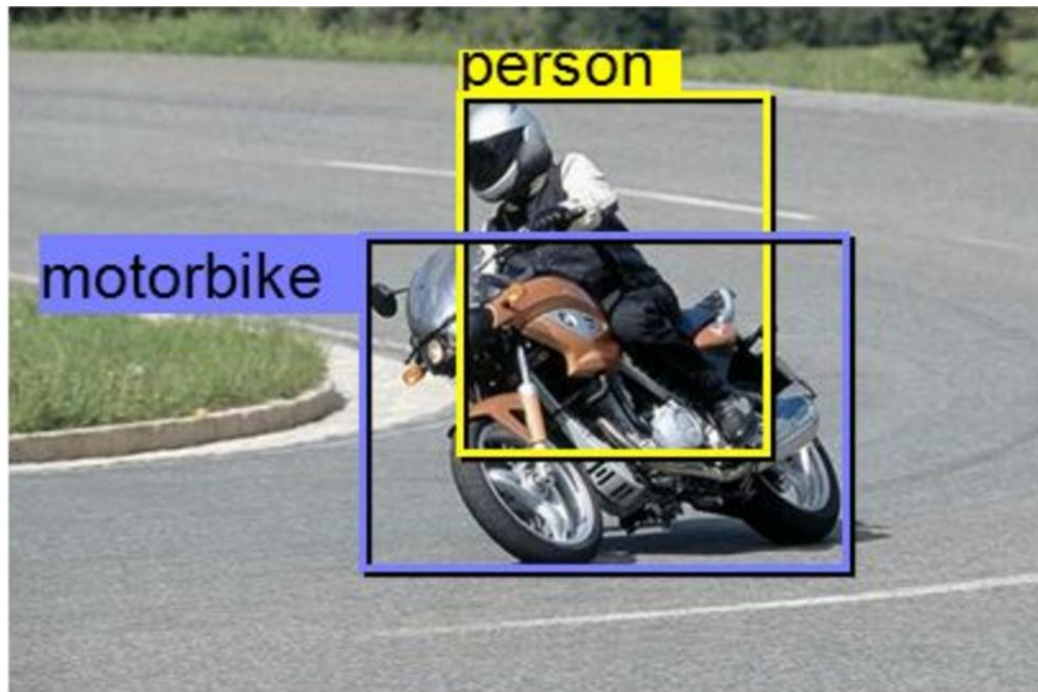
- Person
- Motorcyclist
- Bike



<https://people.eecs.berkeley.edu/~jhoffman/talks/llda-baylearn2014.pdf>

DETECTION

Detect and
label objects



<https://people.eecs.berkeley.edu/~jhoffman/talks/llda-baylearn2014.pdf>

SEMANTIC SEGMENTATION

Label every pixel

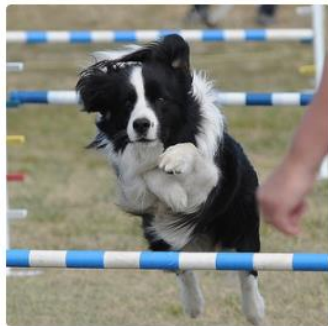


<https://people.eecs.berkeley.edu/~jhoffman/talks/lsta-baylearn2014.pdf>

IMAGE CAPTIONING USING CNN+RNN



"girl in pink dress is jumping in air."



"black and white dog jumps over bar."



"young girl in pink shirt is swinging on swing."



"man in blue wetsuit is surfing on wave."



"little girl is eating piece of cake."



"baseball player is throwing ball in game."



"woman is holding bunch of bananas."



"black cat is sitting on top of suitcase."

Image credit – <http://cs.stanford.edu/people/karpathy/deepimagesent/>

VIDEO CLASSIFICATION & CAPTION USING CNN+RNN



A man is pouring oil into a pot.



A dog is playing in a bowl.



*The person opened the drawer.
The person took out a pot.
The person went to the sink.
The person washed the pot.
The person turned on the stove.*



*The person peeled the fruit.
The person put the fruit in the bowl.
The person sliced the orange.
The person put the pieces in the plate.
The person rinsed the plate in the sink.*

NATURAL LANGUAGE OBJECT RETRIEVAL

a scene with three people query='man far right'



query='man far right'



query='left guy'



query='cyclist'



<http://arxiv.org/pdf/1511.04164v3.pdf>

VISUAL AND TEXTUAL QUESTION ANSWERING



What is the main color on the bus ?



Answer: **blue**



What type of trees are in the background ?



Answer: **pine**



How many pink flags are there ?



Answer: **2**



Is this in the wild ?



Answer: **no**

SPEECH RECOGNITION USING CNN OR RNN

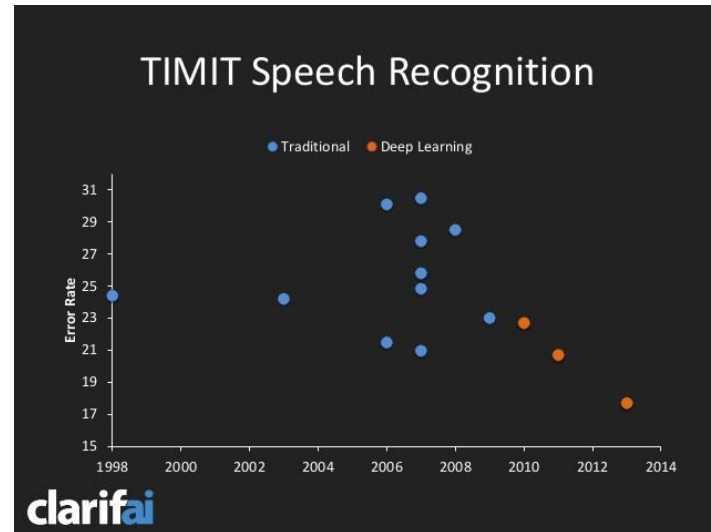
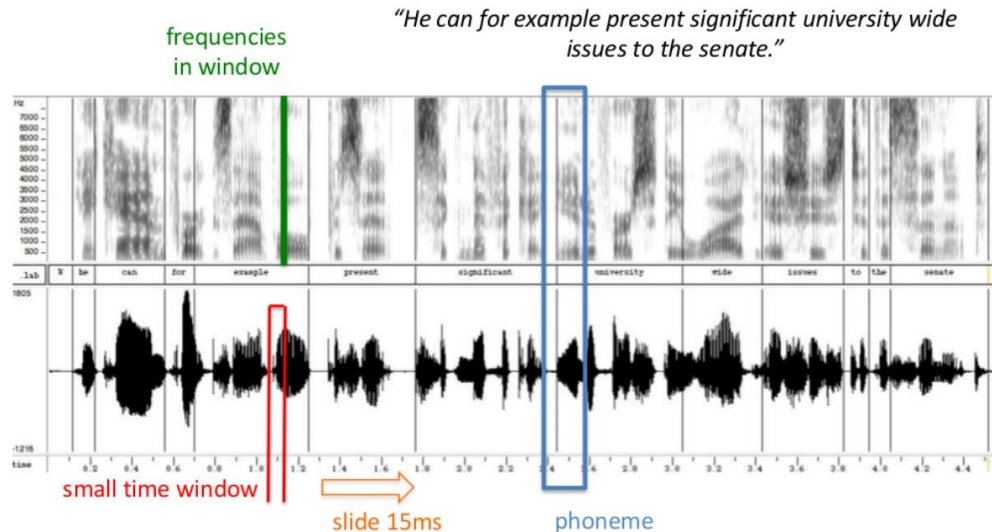


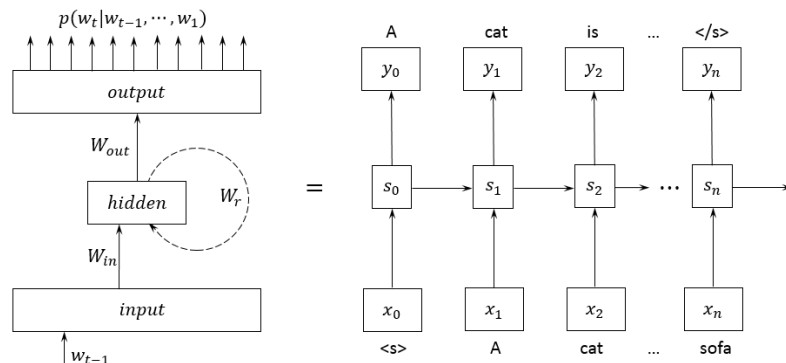
Image credit:

<http://www.slideshare.net/andrewgardner5811/deep-learning-for-data-scientists-dsatl-talk-alpharetta-20140108>

<http://image.slidesharecdn.com/1-141120172105-conversion-gate01/95/clarifai-data-driven-nyc-november-2014-7-638.jpg?cb=1416505093>

NLP / TEXT ANALYTICS : LANGUAGE MODELS

Recurrent Neural Net (RNN) based Language Modeling



$P(w_t | w_0, \dots, w_{t-1})$ Given previous word sequence (history):
predict the next word

Example: <s> A cat is sitting on the sofa </s>

[1] <http://arxiv.org/abs/1312.3005>

[2] Chen, Xie, Wang, Yongqiang, Liu, Xunying, Gales, Mark JF, and Woodland, Philip C. Efficient gpu-based training of recurrent neural network language models using spliced sentence bunch. In INTERSPEECH, 2014

¹ Nvidia Geforce GTX Titan

² Intel Xeon E5-2670 2.6GHz, Intel Compiler ICC 14.0.2

³ Intel Xeon Haswell E5-2697 v3, Red Hat Linux 6.5, Intel Compiler ICPC 16.0.0 20150815, MKL 11.3.0 20150730

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark[®] and MobileMark[®], are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more information go to <http://www.intel.com/performance>

One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling [1]:

*"We show performance of several well-known types of language models, with the best results achieved with a **recurrent neural network** based language model ..."*

GPU Performance claim for RNN based language modeling [2]:

"This (GPU implementation) gives 27 times speed up and a 0.1% absolute reduction in WER over the C-RNNLM baseline"

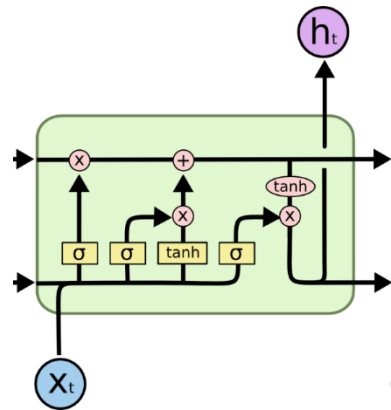
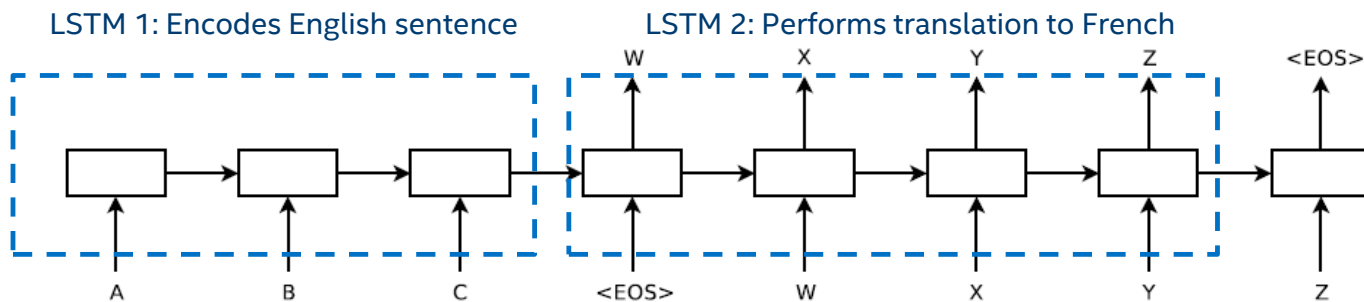
Vocabulary size= **20K** (Google's One Billion Words LM benchmark)

	Throughput (words/sec)		
	Published Result [2]		Our Result
	GPU ¹	CPU ²	CPU ³
RNN 512	9.9k	0.37k	12.6k

NLP / TEXT ANALYTICS : LONG SHORT-TERM MEMORY (LSTM)

Good at learning long term dependencies and is less sensitive to vanishing gradient

- *Machine Translation (Google [1])*
- *Document Classification (Facebook [2])*
- *Context Comprehending (Google/Deepmind [3])*



[1] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. <http://arxiv.org/pdf/1409.3215v3.pdf> in NIPS'14

[2] Xiang Zhang, Junbo Zhao, and Yann LeCun. <http://arxiv.org/pdf/1509.01626v2.pdf> in NIPS'15

[3] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, Phil Blunsom. <http://arxiv.org/pdf/1506.03340v3.pdf> in NIPS'15

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark[®] and MobileMark[™], are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more information go to <http://www.intel.com/performance>

Artificial Intelligence

Sense, learn, reason, act, and adapt to the external world without explicit programming

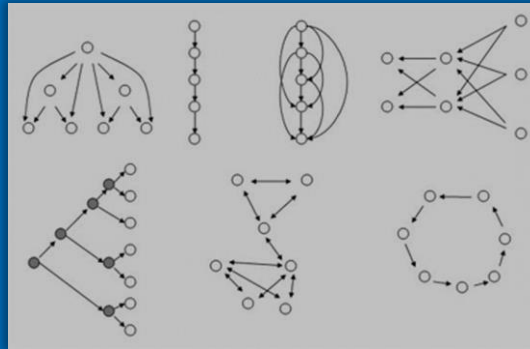
Perception

Detect or recognize patterns in audio, visual, tactile, ambient data



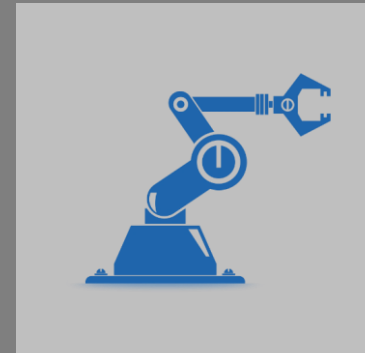
Cognition

Organize data patterns into meaningful structures and recommend action



Action

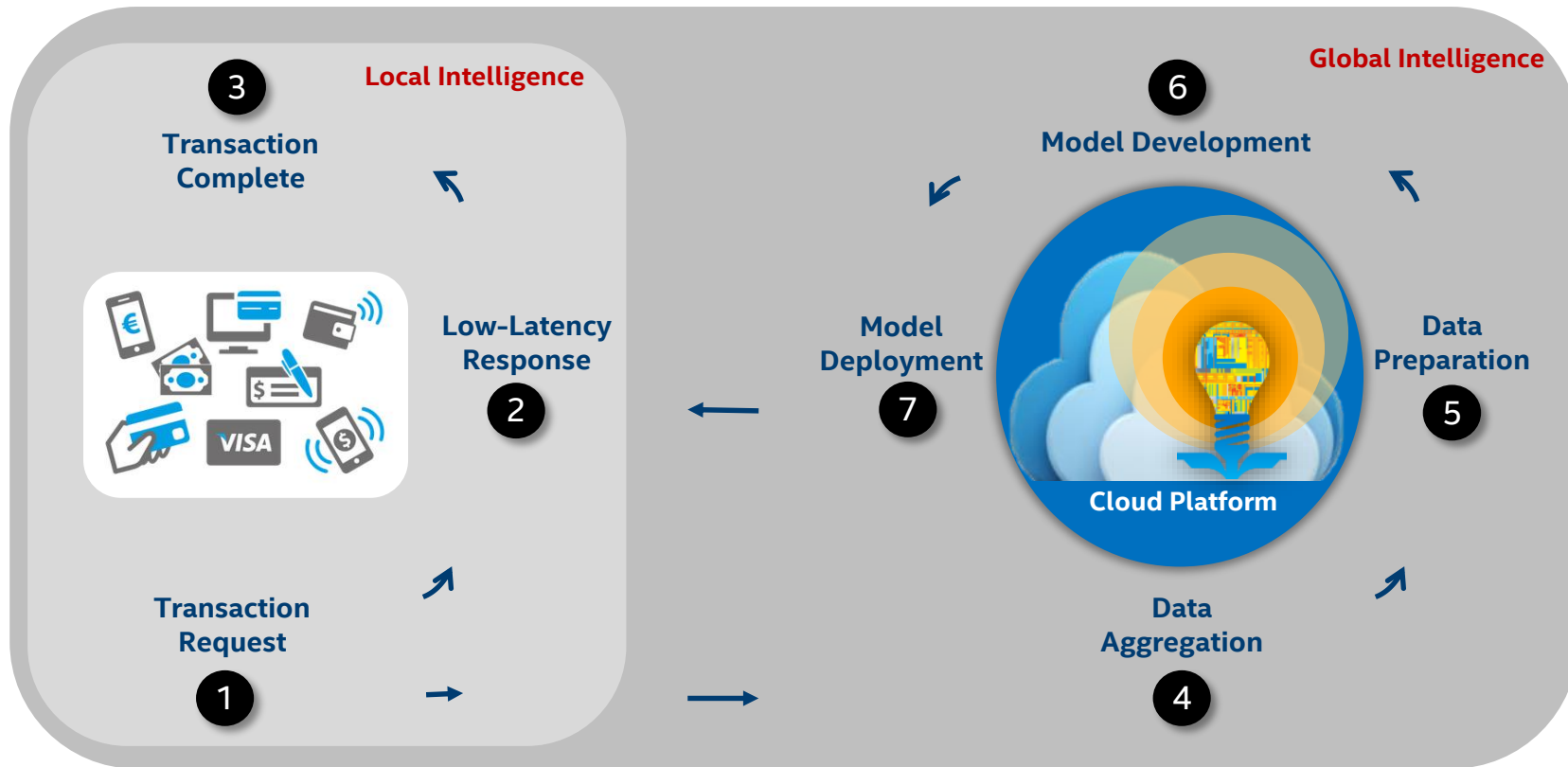
Communicate, control, or respond to external stimuli



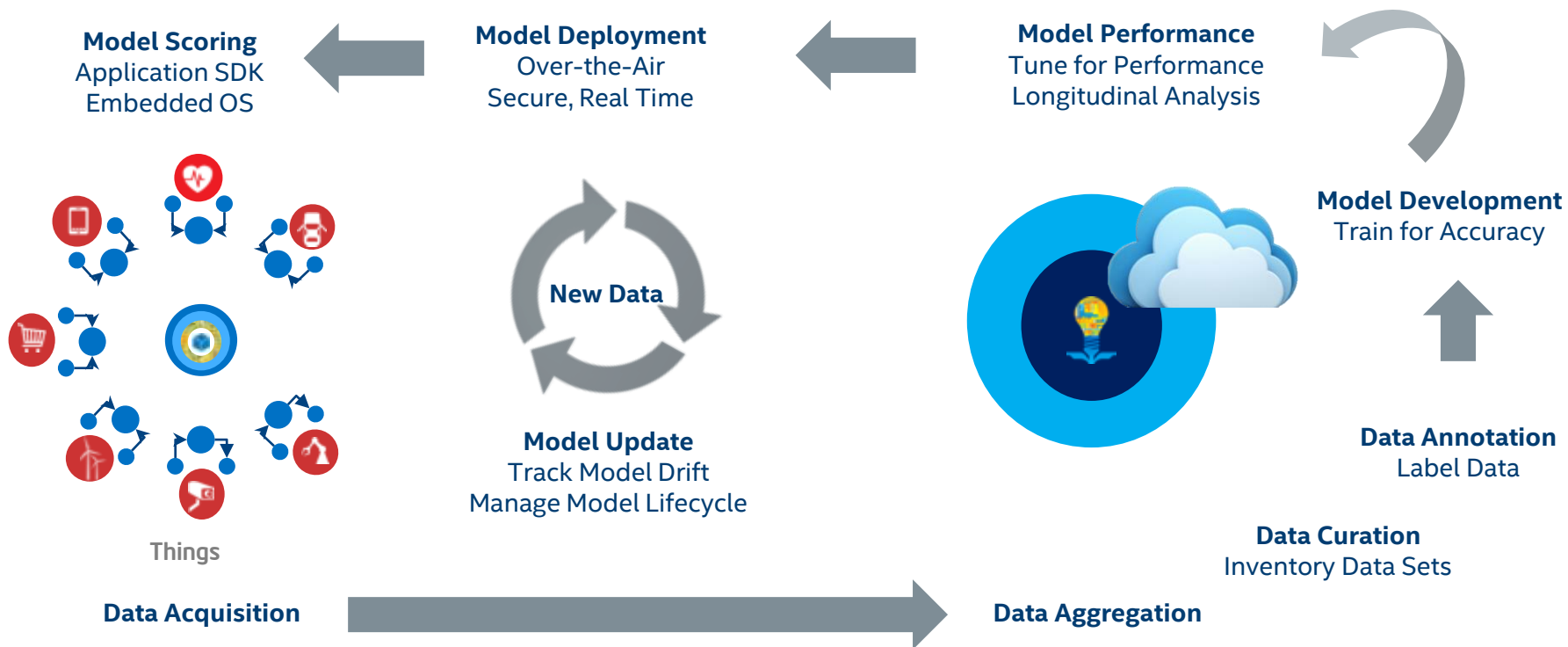
Memory

Store and associate data, patterns, decisions, and actions for recall

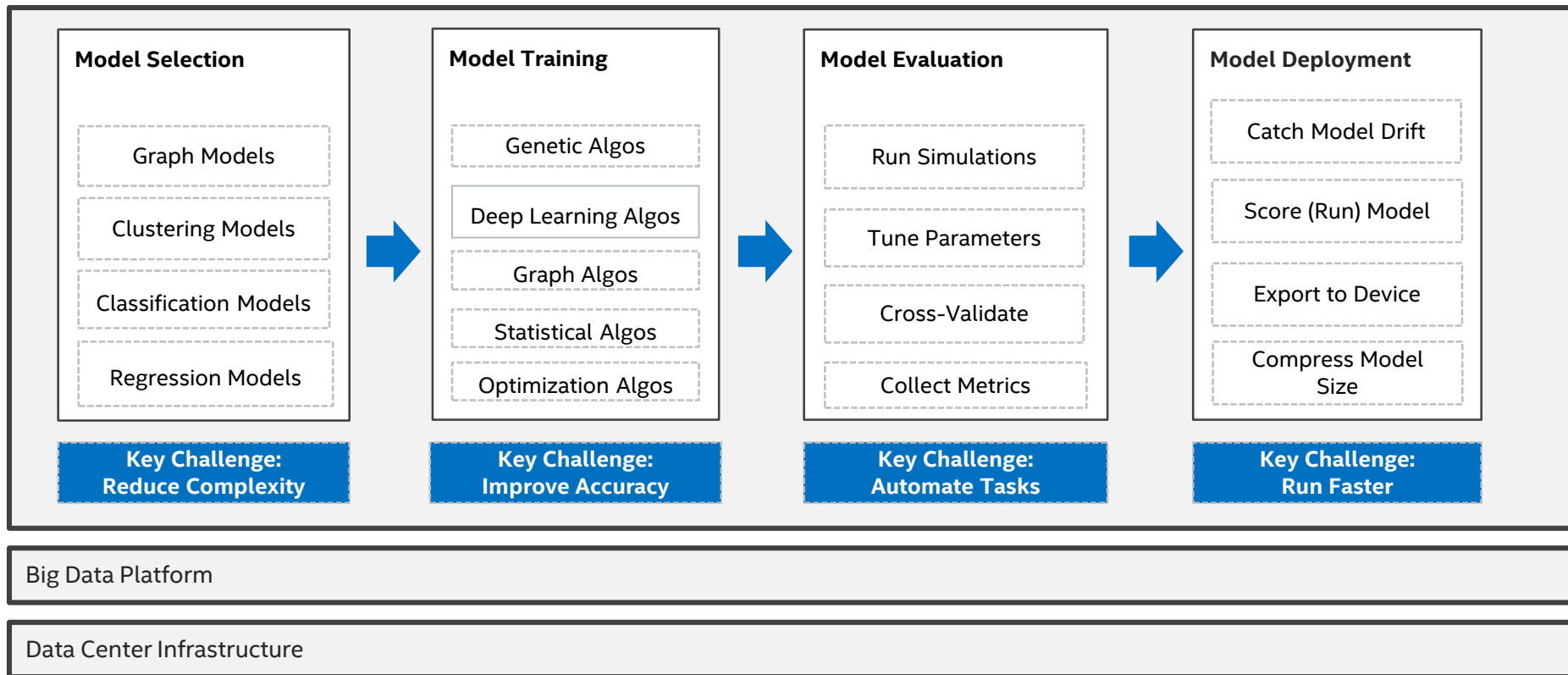
GENERALIZED MODEL-BASED SOLUTION ARCHITECTURE



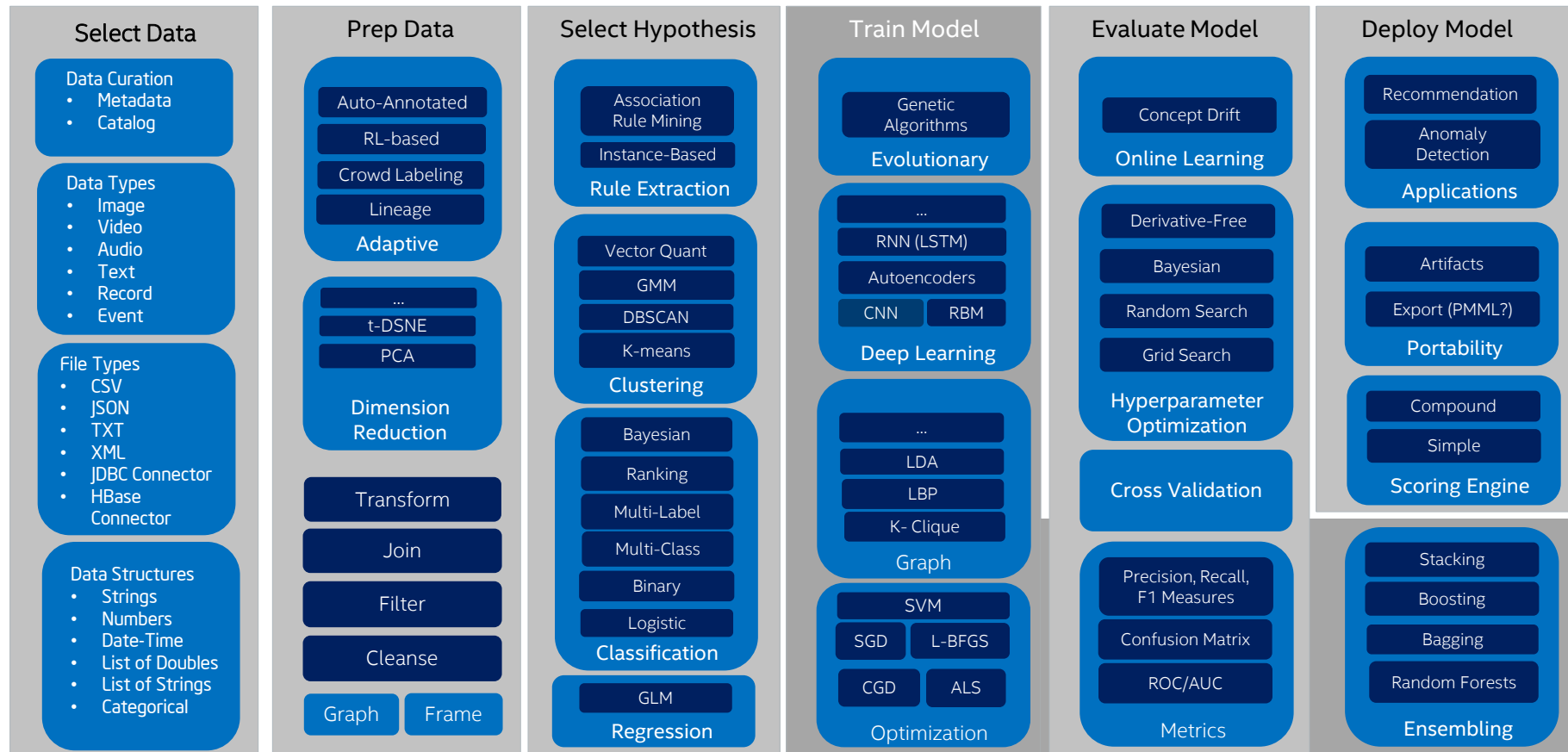
GENERALIZED MODEL-BASED SOLUTION ARCHITECTURE



MODEL DEVELOPMENT: KEY CHALLENGES



MODEL DEVELOPMENT: METHODS & ALGORITHMS



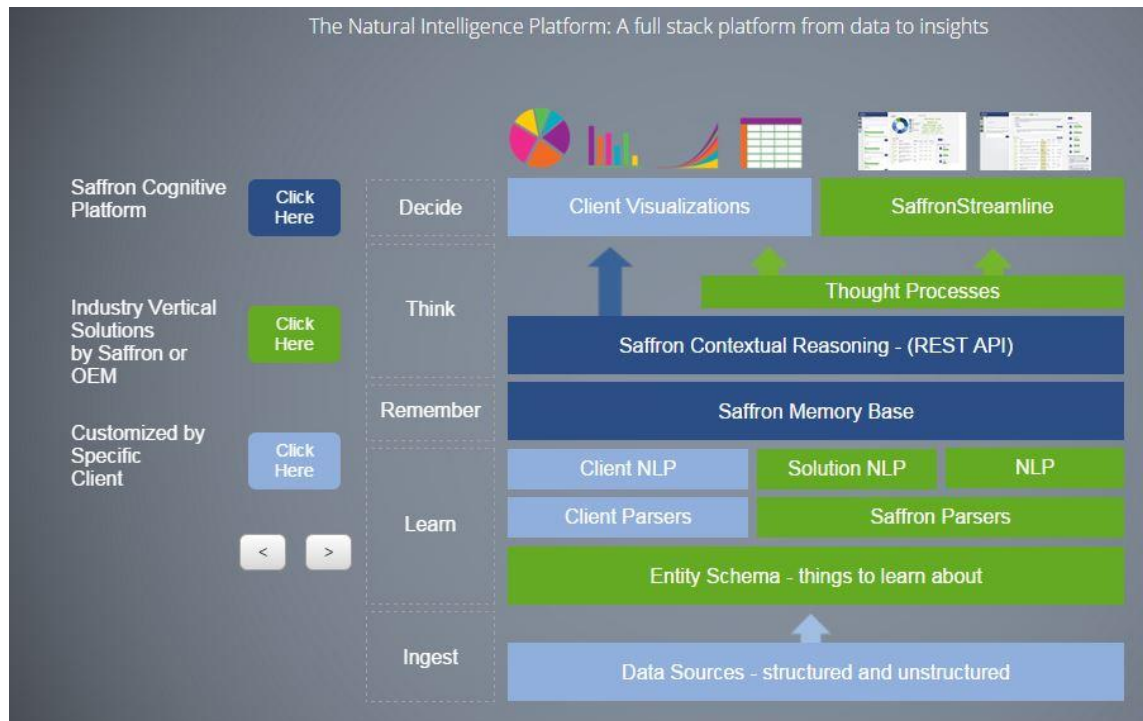
COGNITIVE COMPUTING: SAFFRON & BEYOND

"**Saffron MemoryBase®** is a key:value, incremental learning, fast-query, graph-oriented, matrix-implemented, semantic, and statistical **knowledge store** inspired by the associative structure and function of real neural systems."

<http://saffrontech.com/white-papers/>

Compared to Machine Learning:

- Saffron does not use Models
 - No algorithms
 - No training
 - No scoring
- Saffron does use Graph / Linear Algebra
 - Represents knowledge as a graph of graphs (network of networks)
 - Implements a graph as a matrix
 - Matrix cells hold "associations" (co-occurrence or similarity)
 - Updates matrices as data arrives





- Strength-Based Network Analysis
- Event-based Similarity
- Anomaly Detection
- Explanatory Root Cause
- Sequence And Novelty Detection
- Cognitive Classification
- Temporal Similarity Analysis

Delivered cloud, on
premise or hybrid

AI in the Future

MACHINE LEARNING & DEEP LEARNING FROM INTEL

Solutions



Demonstrate business value of machine learning and deep learning with lighthouse solutions

Trusted Analytics Platform
TAP

Intel® Scalable System Framework

Accelerate adoption of advanced analytics by enabling IoT and cloud-scale analytics platforms



Caffe

theano



Drive standard optimizations across deep learning frameworks using common kernels

Intel® Math Kernel
Library (Intel® MKL)

Intel® Data Analytics Acceleration
Library (Intel® DAAL)

Extract maximum performance from Intel hardware using math kernels and optimized algorithms



Intel® Omni-Path
Architecture (Intel®
OPA)

Optimize single-node and cluster performance on IA while developing new architectures based on parallel computing

INTEL HARDWARE PORTFOLIO FOR DEEP LEARNING

Training



Intel® Xeon Phi™ Processors

- Optimized for performance
- Scales with cluster size for shorter time to model
- x86 architecture, consistent programming model for training and scoring

Inference



Intel® Xeon® Processors

- Optimized for performance/TCO
- Most widely deployed scoring solution



Intel® Xeon® Processors + FPGA (discrete)

- Optimized for performance/watt
- Reconfigurable – can be used to accelerate many DC workloads
- Programmable with industry standard OpenCL



Cost reduction scenarios described are intended as examples of how a given Intel- based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction. Results have been estimated or simulated using internal Intel analysis or architecture simulation or modeling, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance.

INTEL® XEON PHI™ PROCESSOR



No PCIe Dependency

Bootable host processor

Topple Memory Wall

Integrated memory up to 16GB

Run Any Workload

Intel® Xeon® processor binary-compatible

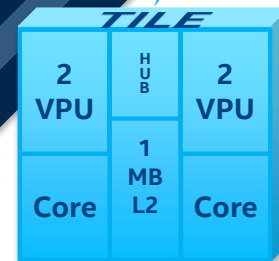
Bootable Host CPU

Integrated Memory

Platform Memory (DDR4)

Integrated Fabric

Processor Package



Scale Out Seamlessly

Bootable CPU, element of Intel® SSF

Reduce Cost

Dual-port Intel® Omni-Path Fabric

Raise Memory Ceiling

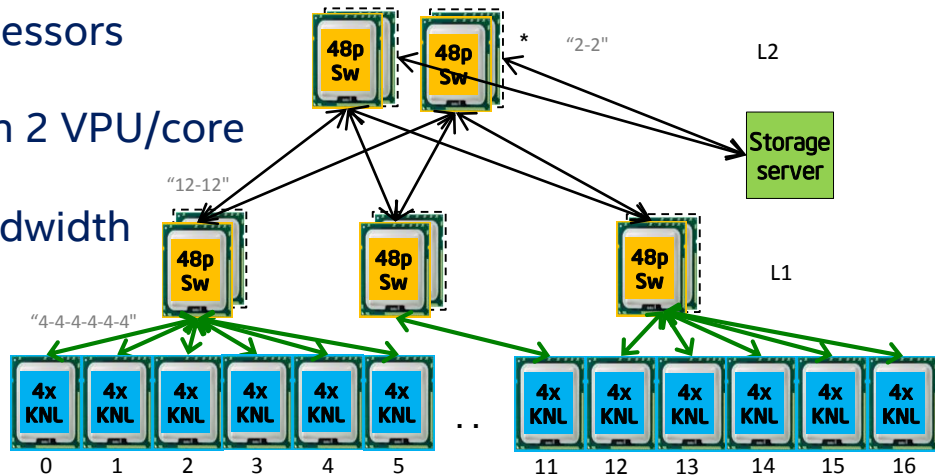
Platform memory up to 384 GB (DDR4)

INTEL® XEON PHI™ + OMNI-PATH

Deep Learning Training HW

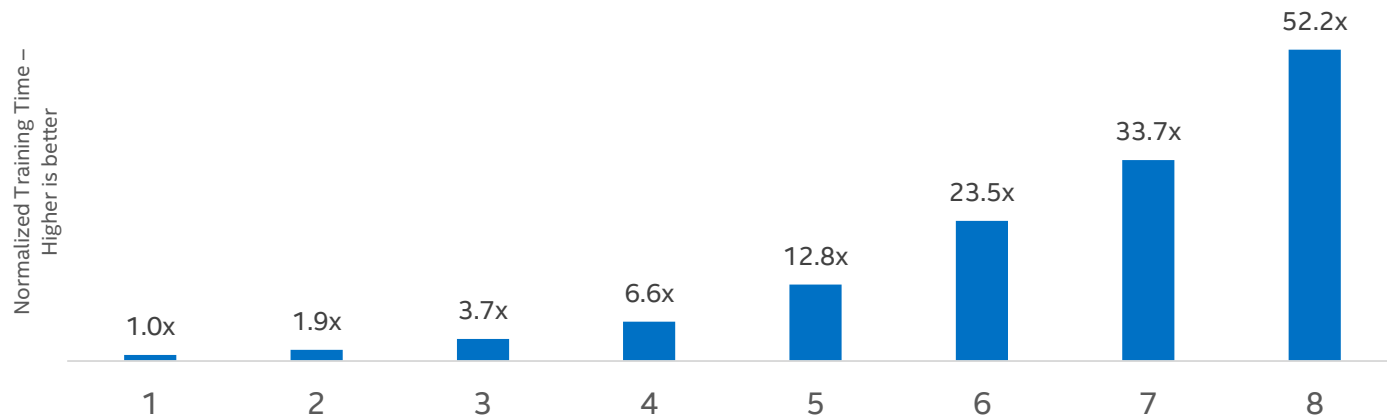
Knights Landing

- Up to >6 peak SP TFLOPs per socket
- Binary-compatible with Intel® Xeon® processors
- Up to 72 cores, 512-bit SIMD vectors with 2 VPU/core
- Integrated memory delivers superior bandwidth
- Integrated Intel® Omni-Path fabric (dual-port; 50 Gb/s ↔)
- **Distributes the training workload**



TRAIN UP TO 50X FASTER WITH INTEL® XEON PHI™ PROCESSOR

Deep Learning Image Classification Training Performance - MULTI-NODE Scaling



Topology: AlexNet*

Dataset: Large image database

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more information go to <http://www.intel.com/performance/datacenter>. Configurations: Up to 50x faster training on 128-node as compared to single-node based on AlexNet* topology workload (batch size = 1024) training time using a large image database running one node Intel Xeon Phi processor 7250 (16 GB MCDRAM, 1.4 GHz, 68 Cores) in Intel® Server System LADMP2312KXXX41, 96GB DDR4-2400 MHz, quad cluster mode, MCDRAM flat memory mode, Red Hat Enterprise Linux* 6.7 (Santiago), 1.0 TB SATA drive WD1003FZEX-00MK2A0 System Disk, running Intel® Optimized DNN Framework, training in 39.17 hours compared to 128-node identically configured with Intel® Omni-Path Host Fabric Interface Adapter 100 Series 1 Port PCIe x16 connectors training in 0.75 hours. Contact your Intel representative for more information on how to obtain the binary. For information on workload, see <https://papers.nips.cc/paper/4824-Large-image-database-classification-with-deep-convolutional-neural-networks.pdf>.

INTEL SW DEVELOPMENT FOR DEEP LEARNING

Trusted Analytics Platform



Open Source Frameworks



Intel® Math Kernel Library (Intel® MKL)



Intel® Data Analytics Acceleration Library (Intel® DAAL)



Overview	Single platform from Data Science to Application Development	Toolkits driven by Academia and Industry for scripting and training ML algorithms	High Performance Math Primitives granting low level of control	Broad Data Analytics Acceleration object oriented library supporting distributed ML at the algorithm level
Audience	Application Developers and Data Scientists	Machine Learning Researchers and Data Scientists	Consumed by developers of higher level libraries and Applications	Wider Data Analytics and ML audience, Algorithm level development, needing predesigned algorithms for all stages of data analytics
Example Usage	Application creation from the Big Data infrastructure, Data Science tools all the way to app development	Script and Train a Convolution Neural Network for Image Recognition in Python*	Call Matrix Multiplication, Convolution Functions	Call K-Means, Linear Regression, ALS Algorithms

INTEL® MATH KERNEL LIBRARY (INTEL® MKL)

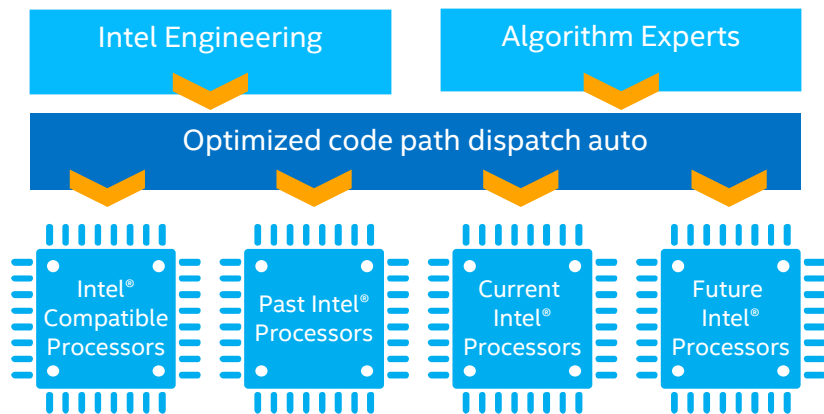
Highly optimized threaded math routines

- Performance, Performance, Performance!

Industry's leading math library

- Widely used in science, engineering, data processing

Tuned for Intel® processors – current and next generation



EDC North America
Development Survey
2011, Volume II

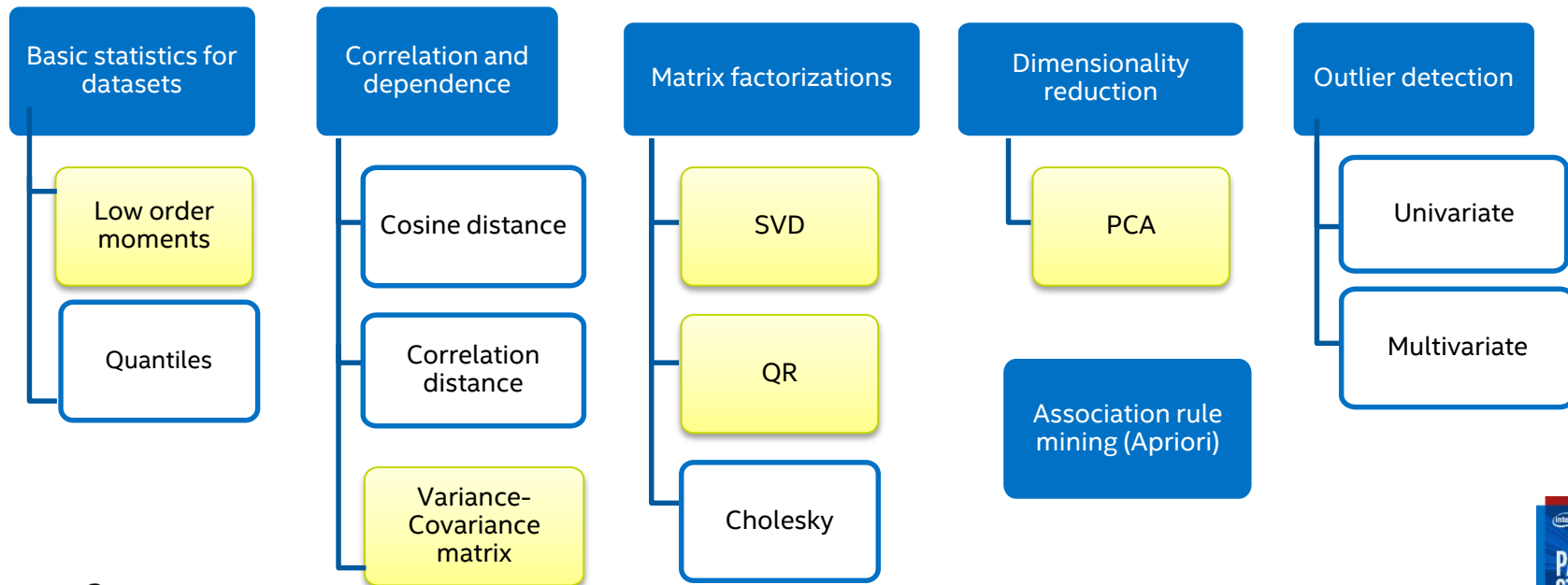
33% of math libraries users rely on
Intel® Math Kernel Library


Be multiprocessor aware

- Cross-Platform Support
- Be vectorised, threaded, and distributed multiprocessor aware

INTEL® DATA ANALYTICS ACCELERATION LIBRARY (INTEL® DAAL)

Provides building blocks for all data analytics stages, from data preparation to data mining & machine learning

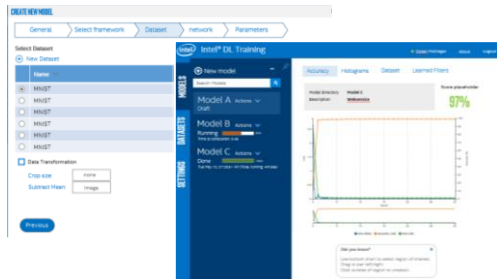


 Algorithms supporting streaming and distributed processing



INTEL® DEEP LEARNING SDK

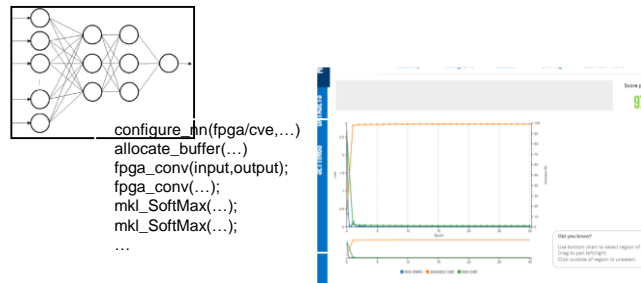
Intel Deep Learning Training Tool



- Simplify installation of Intel optimized Deep Learning Frameworks
- Easy and Visual way to Set-up, Tune and Run Deep Learning Algorithms:

MKL-DNN Optimized
Deep Learning Frameworks

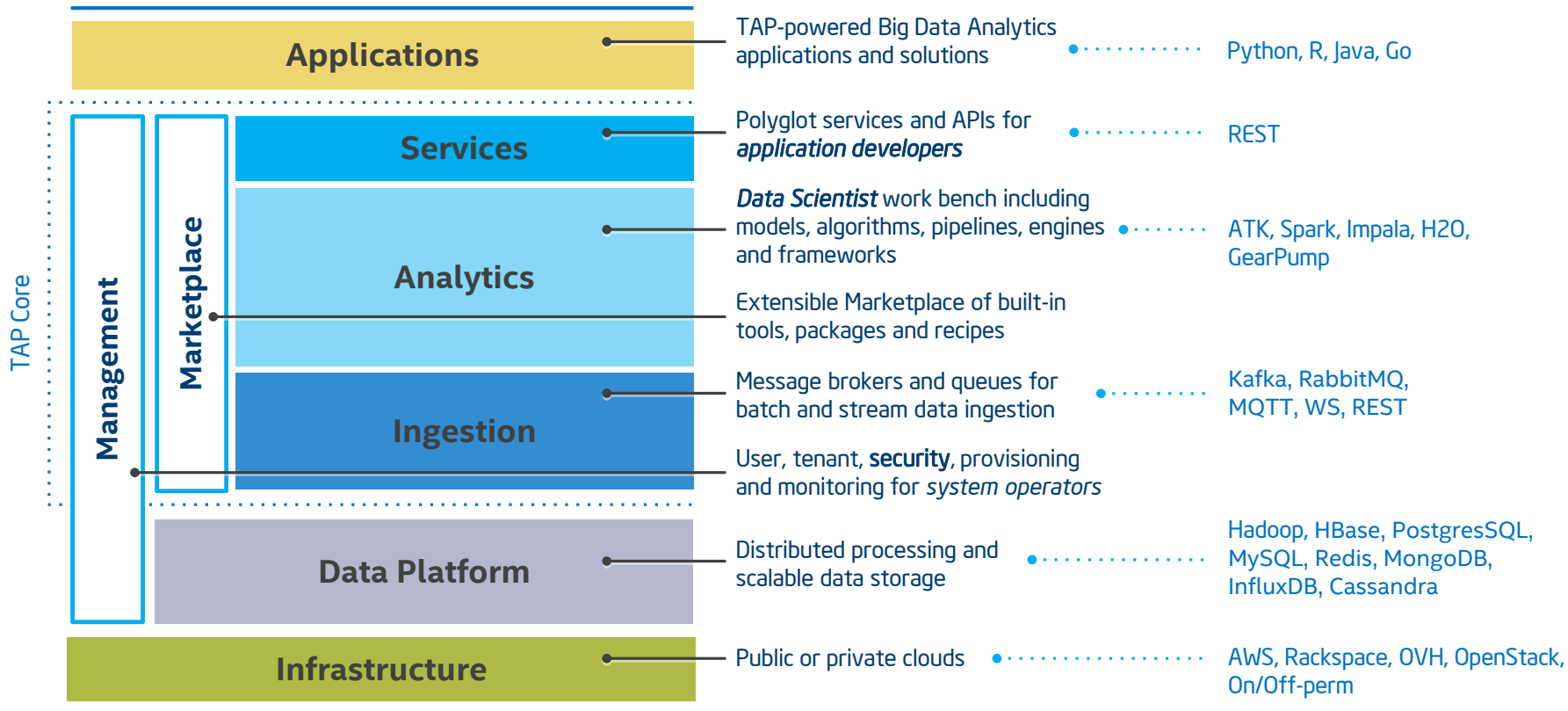
Intel Deep Learning Deployment Tool



- Unlock fast scoring performance on Intel products while abstracting the HW from developers

Optimized libraries & run-times
(MKL-DNN, OpenVX, OpenCL)

TRUSTED ANALYTICS PLATFORM



RESOURCES

<http://www.intel.com/machinelearning>

Legal Notices & Disclaimers

This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Learn more at intel.com, or from the OEM or retailer. No computer system can be absolutely secure.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit <http://www.intel.com/performance>.

Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

Statements in this document that refer to Intel's plans and expectations for the quarter, the year, and the future, are forward-looking statements that involve a number of risks and uncertainties. A detailed discussion of the factors that could affect Intel's results and plans is included in Intel's SEC filings, including the annual report on Form 10-K.

The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

Intel, the Intel logo, Pentium, Celeron, Atom, Core, Xeon and others are trademarks of Intel Corporation in the U.S. and/or other countries.

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

© 2016 Intel Corporation.

