Virtuous Cycle of AI

Pradeep K Dubey

Intel Senior Fellow and Fellow of IEEE

Director Parallel Computing Lab, Intel Labs

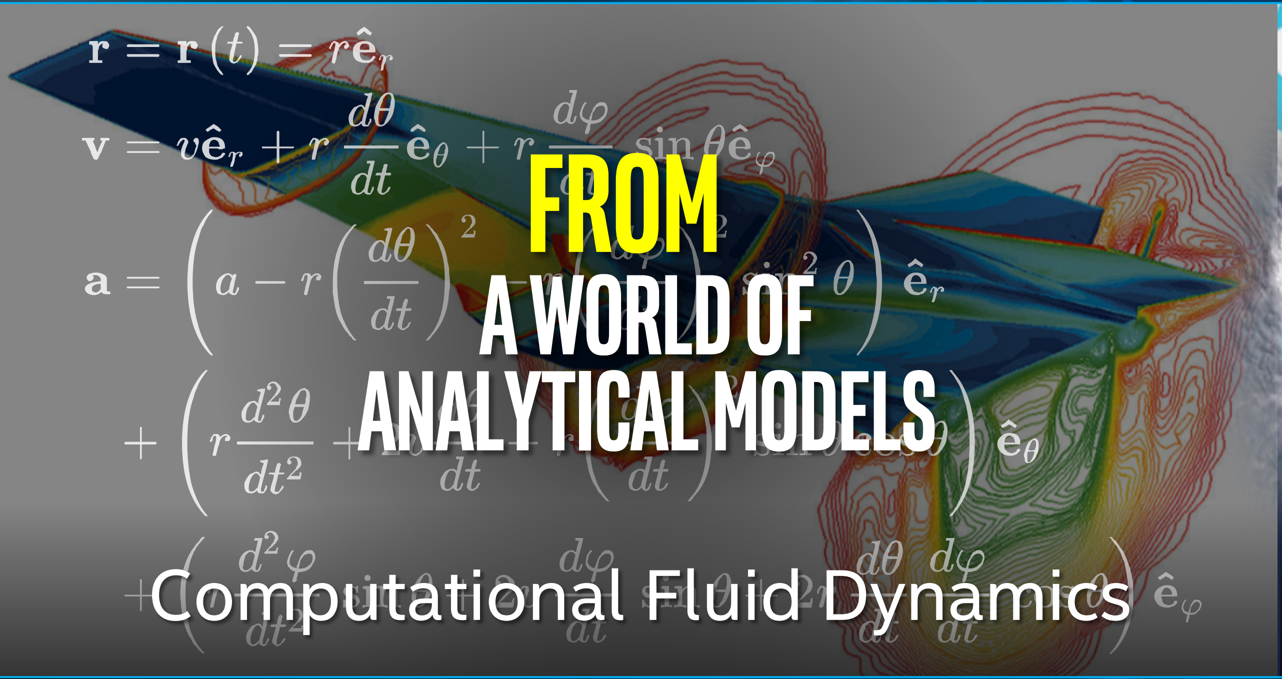
3rd Annual Meeting of Heterogenous Integration Roadmap (IEEE EPS Chapter)
Milpitas, California Feb 20, 2020

Machines:
Number Crunching
AND
Decision Making

Division of Labor Between Man and Machine Is Getting Disrupted:
Faster than Anyone Predicted!

The new frontier

Inside - Out ⇄ *Outside - In*



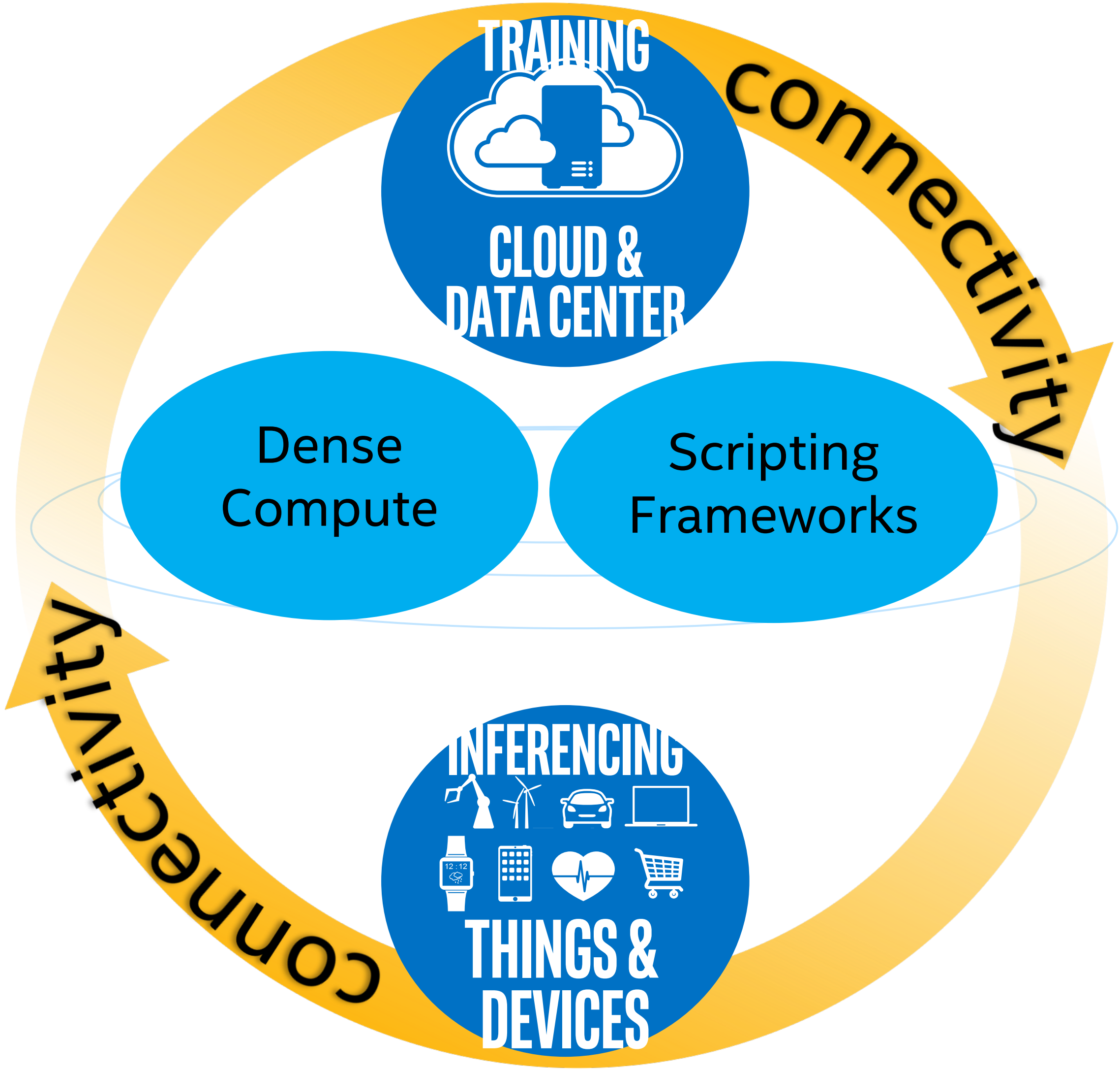
Start with Mathematical Model

Model → Simulate → Predict

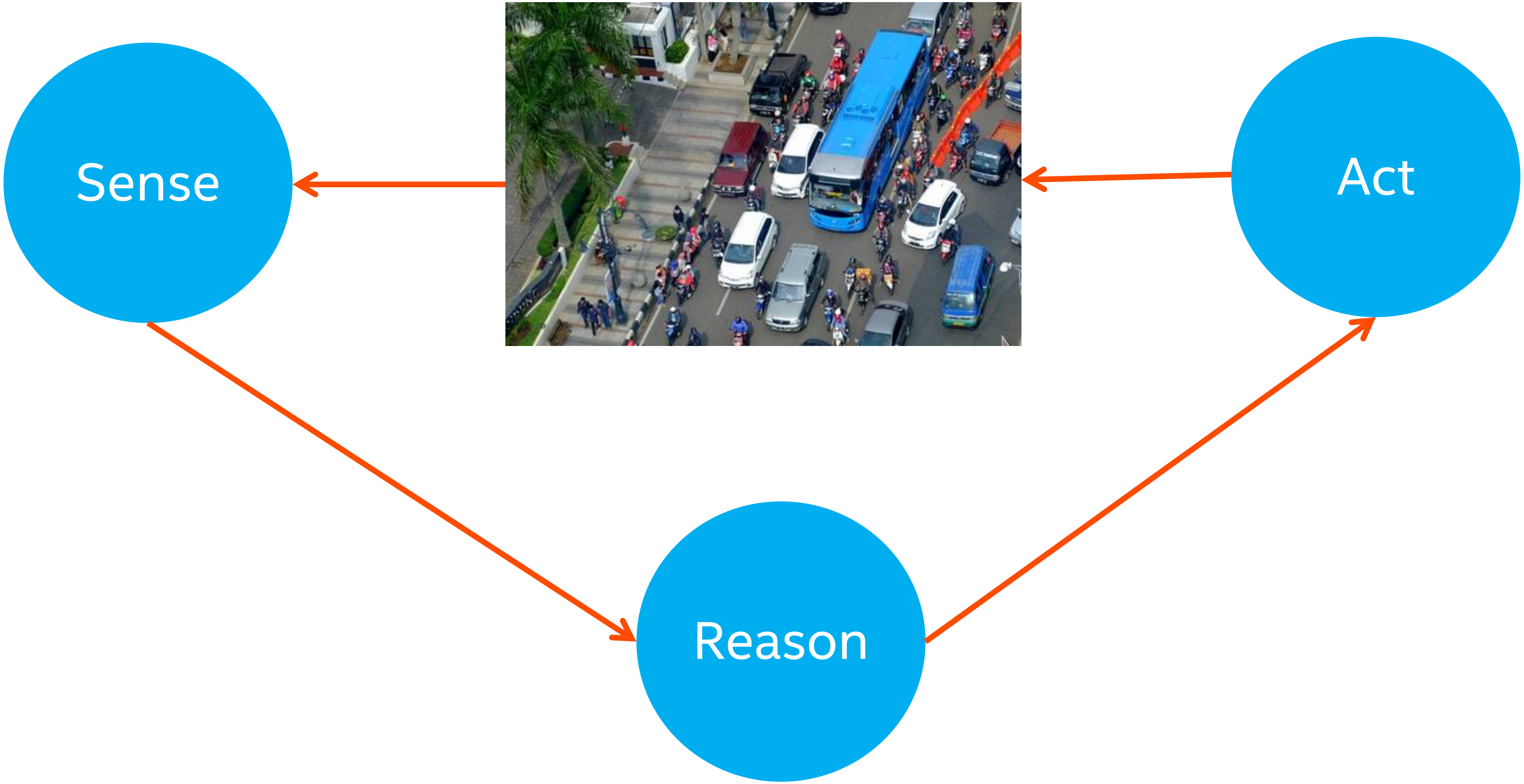
Start with Data

Initial State → Increment → Steer

Virtuous Demand Cycle of AI Compute

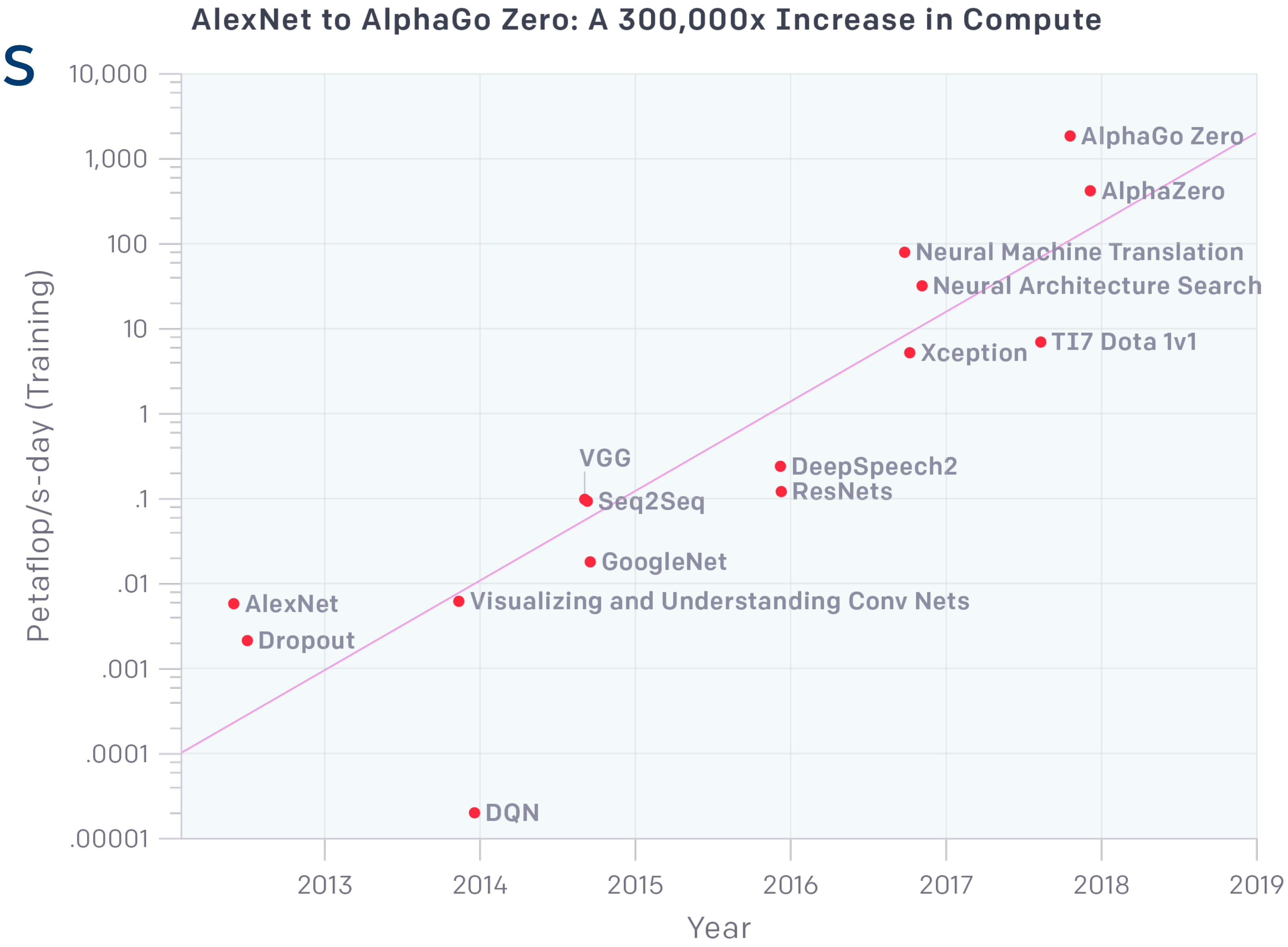


Virtuous Demand Cycle of Compute: A Functional View



AI Compute Needs

AlphaGo Zero needs:
2 EF-days to train →
Need a 100 ExaFlop machine
to train within an hour *

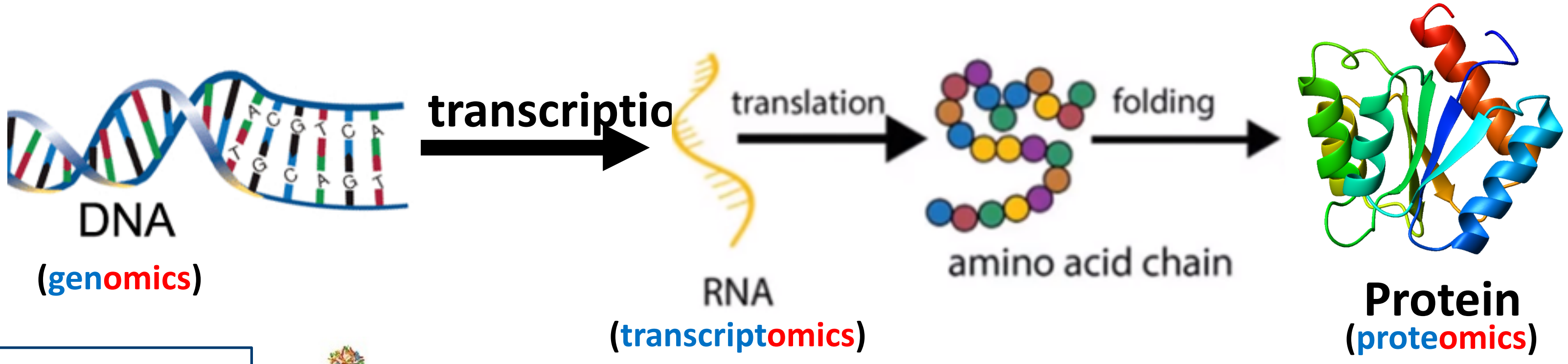


* https://blog.openai.com/ai-and-compute/

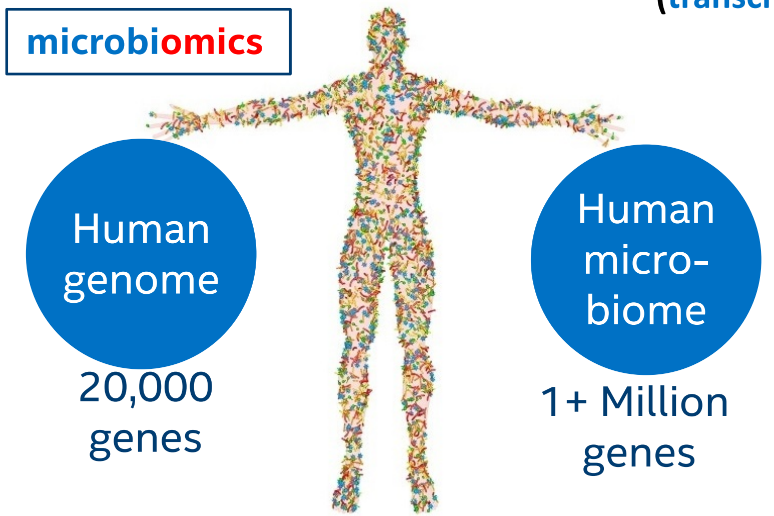
Illustrating population-scale AI in health sector

Explosion of trans-OMICS data in the coming decade

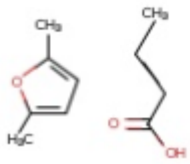
- From the molecular level to the genomic level to the organ level
- Interactions with medications, nutrients, therapeutic devices, and the environment.



microbiomics



metabolomics



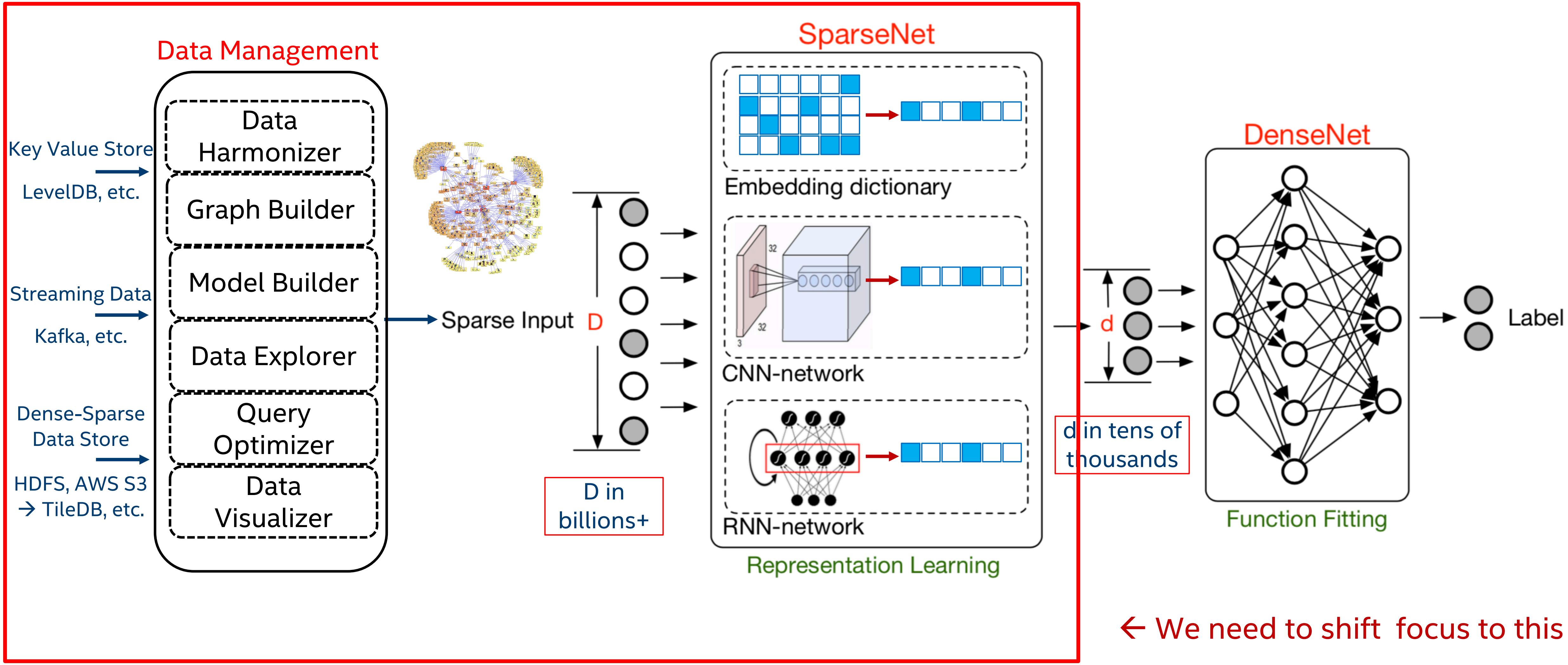
Concentration of various molecules

Radiology
Digital histopathology
Electronic Health Records
…

Platform Characteristics

- **High throughput** omics pipeline over O(Billion) record data sets
- A multi-modal **AI pipeline** to correlate omics data across multiple datasets & generates actionable recommendations
- An open, modular **data-centric-computing platform** optimized for security, scalability and performance across many workloads
- 100s of exabytes of **storage capacity, processed in minutes** for results
- Best-in-class TCO with **easy to use/extend SW-Stack**

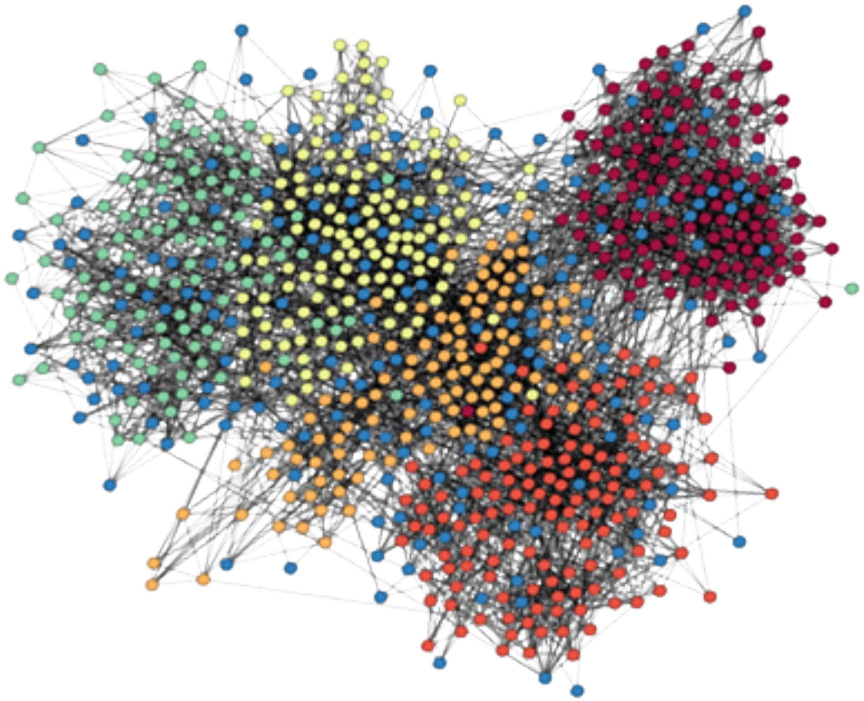
High-dimensional learning Infrastructure



← We need to shift focus to this

Graph Analytics

Improve social-network analysis, fraud-ring detection, anomaly detection



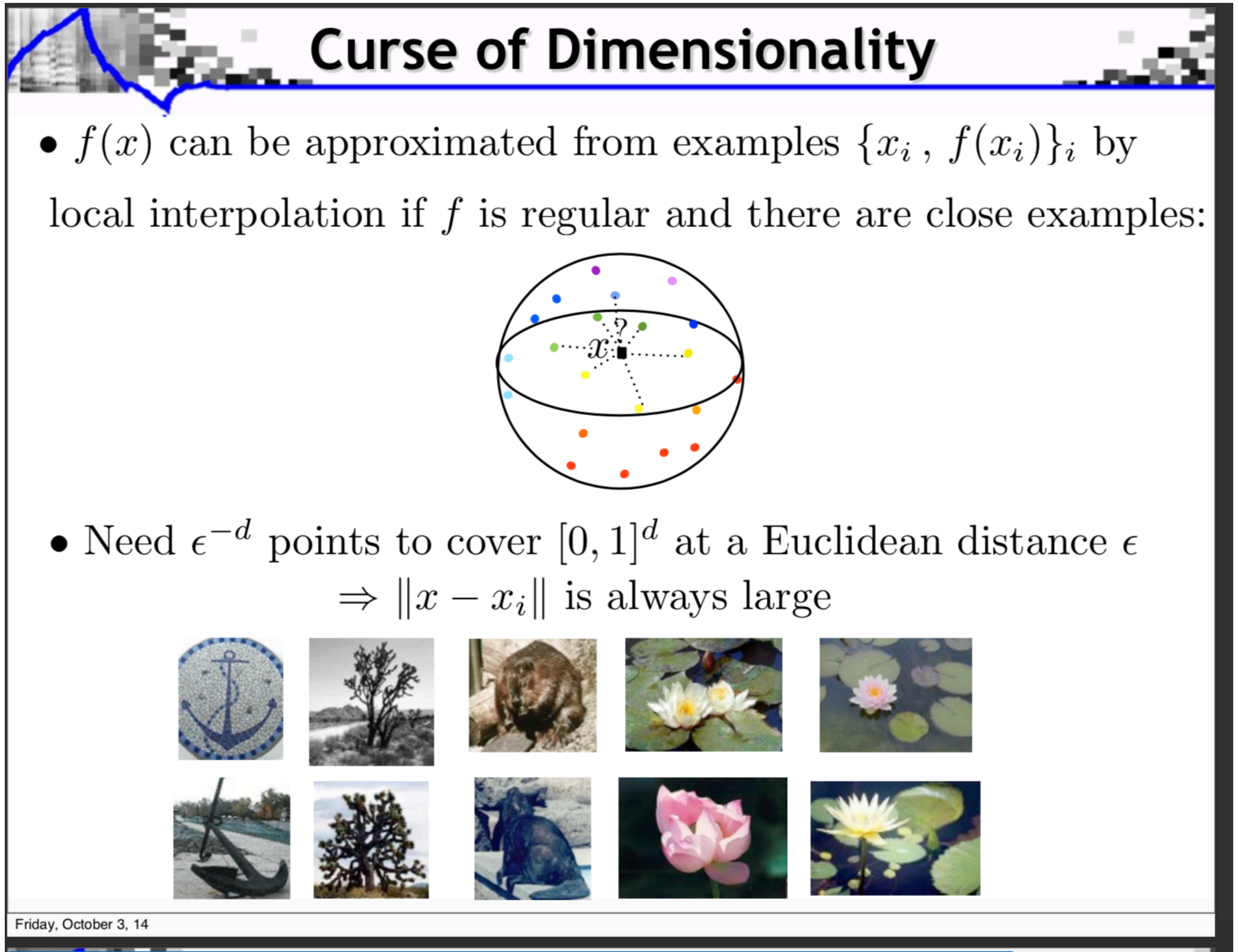
CHALLENGES

- Sparse and irregular memory accesses
- Small data accesses with frequent synchronization
- Scaling to very large datasets

Real-Time Decision Making

Simplified ML Theory World View

- If we only solve a **single** problem and design world-class hardware-software stack for it, what would that be:
  - *High Dimensional Learning*
  - *"Learn a super-compact,deep (hierarchical) approximation of dynamic graphs – computable in polynomial time, and evolving very slowly in time"*

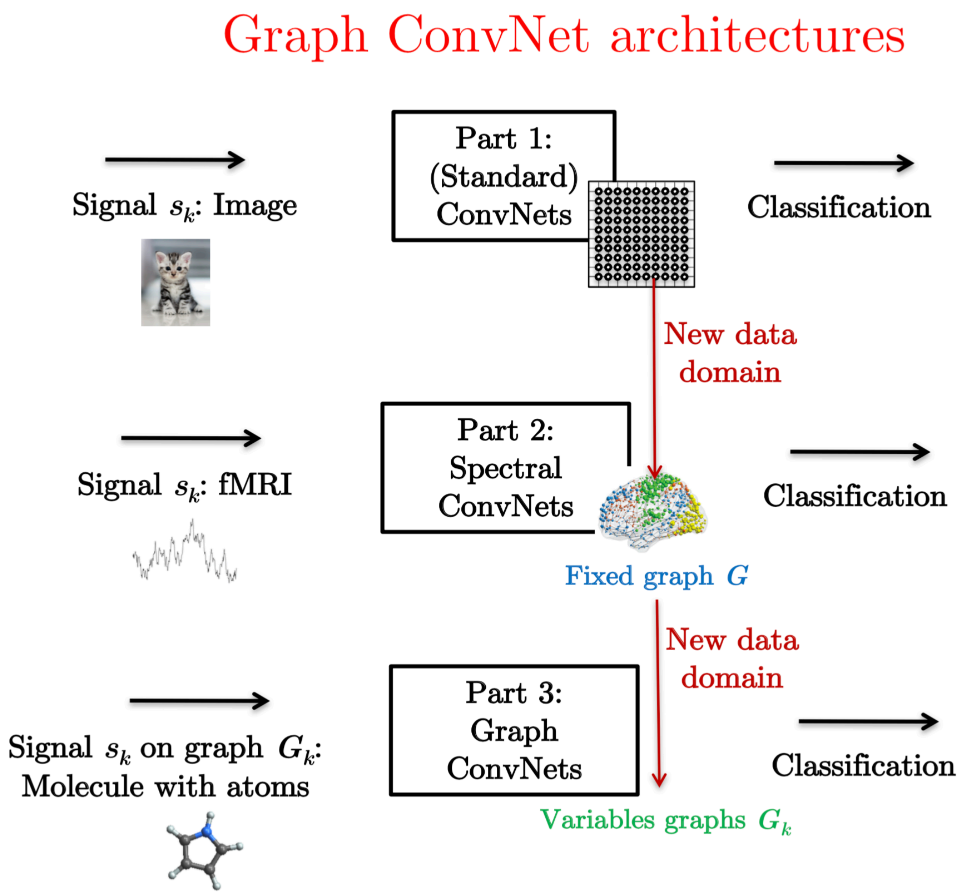


Credit for this slide above goes to Prof. Stephane Mallat *

Recent (2017) Algorithmic Breakthrough: Geometric Deep Learning **

* High Dimensional Learning and Deep Neural Networks: https://www.youtube.com/watch?v=nHXO43BqeQw
** Geometric Deep Learning: Going beyond Euclidean data, by Michael Bronstein, John Bruna, Yann LeCunn, Arthur Sxlam, Pierre Vandergheyn: https://arxiv.org/pdf/1611.08097.pdf

Will convolutions + downsampling prove effective in non-Euclidiean space as well?



IPAM 2018 Workshop on New Deep Learning: https://www.ipam.ucla.edu/programs/workshops/new-deep-learning-techniques/?tab=overview

DARPA Graph Analytics Challenge

DARPA Taps Intel for Graph Analytics Chip Project

Michael Feldman | June 7, 2017 04:22 CEST

Intel shook hands with DARPA to craft HIVE Big Data platform a reality

SR Siliconreview Team
2017-06-09

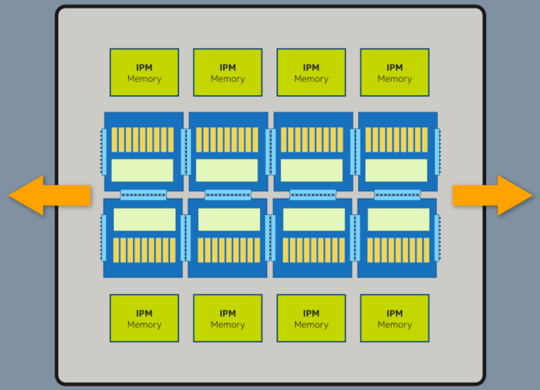
INTEL NAMED TO DARPA PROJECT FOCUSED ON MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

Intel has been selected by DARPA, a U.S. Department of Defense agency, to collaborate on the development of a powerful new data-handling and computing platform that will leverage machine learning and other artificial intelligence (AI) techniques.

PROGRAM TO DEVELOP NEW TECHNOLOGIES TO REALIZE 1,000X PERFORMANCE-PER-WATT GAINS IN THE ABILITY TO HANDLE GRAPH ANALYTICS

Key Technologies and Scalability

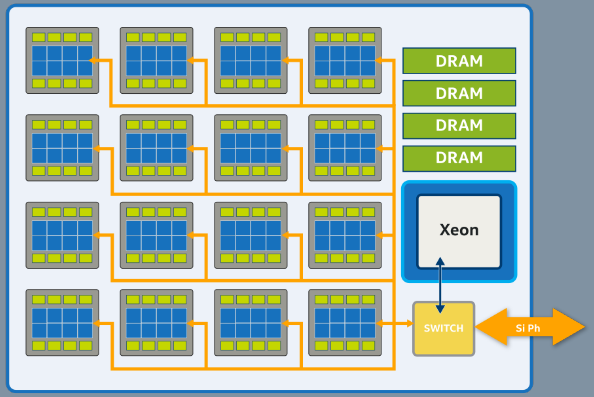
RE-IMAGINED ARCHITECTURE



CPU Support for Small, Irregular Memory Accesses

Near-Memory Atomics

FULLY INTEGRATED



Global Memory Model

Packaging for High I/O and Memory bandwidth

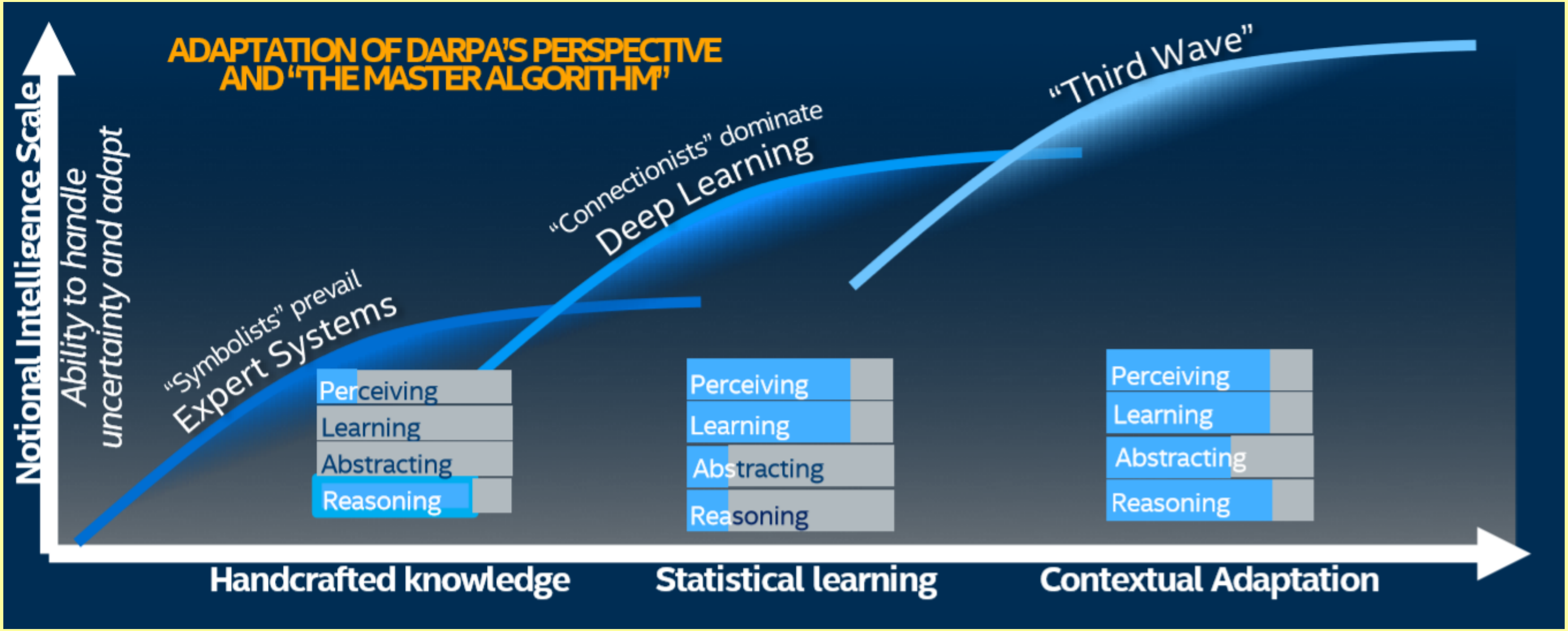
SEAMLESS SCALING

Network as 1st –class Citizen

Flatten Latency Hierarchy

Point-to-point Messages

The Future: Third Wave of AI



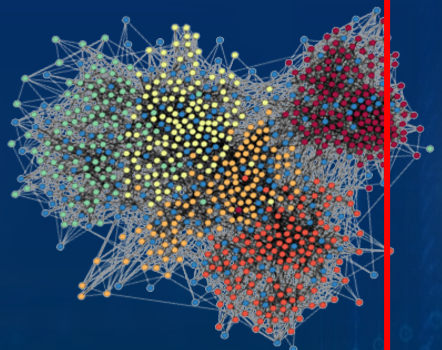
Deep Neural Networks Getting Augmented: NN + X + Memory
Such As: CNN + Bayes Net + Sparse Embeddings

LOOKING AHEAD …

CONVENTIONAL
- Known procedures
- Generate answers

DEEP LEARNING
- Known answers
- Generate procedures with training

GRAPH ANALYTICS
- Graph edges and vertices represent relationships
- Big, sparse data structures



NEUROMORPHIC
- Many procedures
- Adapt the answer with reinforcement

QUANTUM
- Answers superimposed
- Select and measure the answer

What makes delivering AI hard and fun!

Better model building

LEARNING WITH LESS DATA AND SUPERVISION

DEEP NEURAL NETWORKS GETTING AUGMENTED: DATA-DRIVEN + ANALYTICAL + MEMORY

LEARNING MODELS THAT ARE EASIER TO REASON AND BETTER TO GENERALIZE

CONTINUOUS LEARNING FOR MISSION-CRITICAL AI

More efficient and pervasive model deployment

THROUGHPUT, ACCURACY, AND MODEL SIZE TRADEOFFS: SPARSIFICATION AND PRUNING

SELF-LEARNING AND PERSONALIZATION AT THE EDGE

Compute architecture needs of AI

REDUCING ARITHMETIC PRECISION WHILE PRESERVING ACCURACY

FEEDING THE COMPUTE ← MEMORY AND NETWORK; COMPUTE NEAR NETWORK

DOMAIN-SPECIFIC ARCHITECTURES → TRADITIONAL, SPATIAL, NEUROMORPHIC, QUANTUM

Productivity and Scaling needs of AI

STRONG-SCALING AI TO HPC SCALE ON CLOUD INFRASTURCTURE: HIGHER-ORDER METHODS

DELIVERING PERFORMANCE-PRODUCTIVITY: FaaS AND NEW LANGUAGES AND ABSTRACTIONS

We have a problem though …

Where Are All the Ninjas?

2019-01-17 / DEVELOPMENT

The Talent Shortage of Software Developers in 2019

1. According to Code.org, there are less than 50,000 Computer Science graduates in 2017. But, there are over 500,000 open computing positions in the United States. This could mean that in 2020, the available seats for this position will exceed qualified applicants by a million which could widen the gap even more.

Only 10% of programmer pool have academic training to be *ninjas*
New additions to programing pool are data scientists
Hardware is getting increasingly complex for extracting high performance
Hence, *ninja gap* is not likely to go down!

https://fullscale.io/the-talent-shortage-of-software-developers-in-2019/

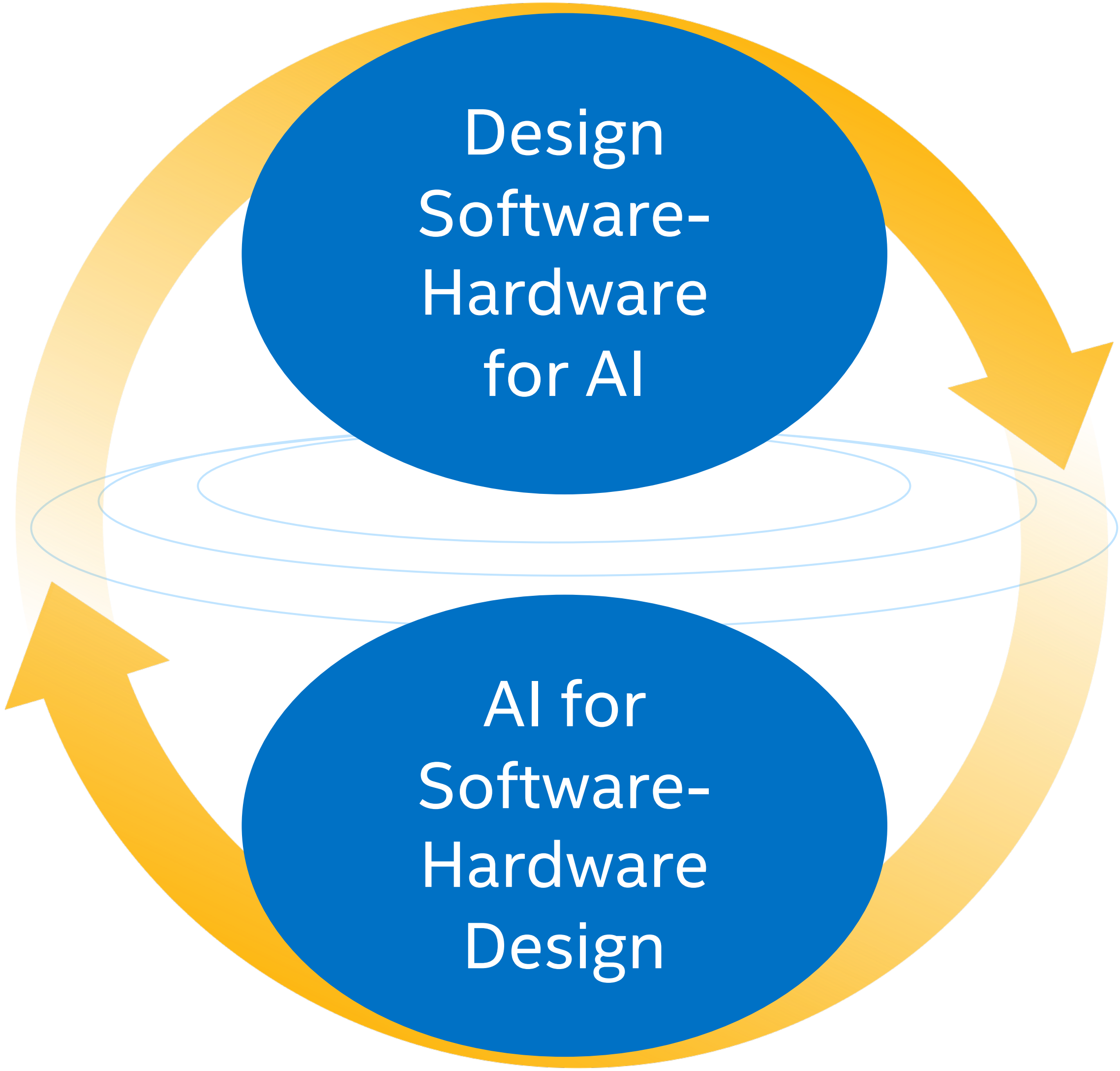
Time to switch gear ….

From:
Designing hardware/software to meet AI needs

To:
AI helping us with our software-hardware needs

This implies a virtuous cycle for realized AI performance growth!

Virtuous Supply Cycle of AI Compute



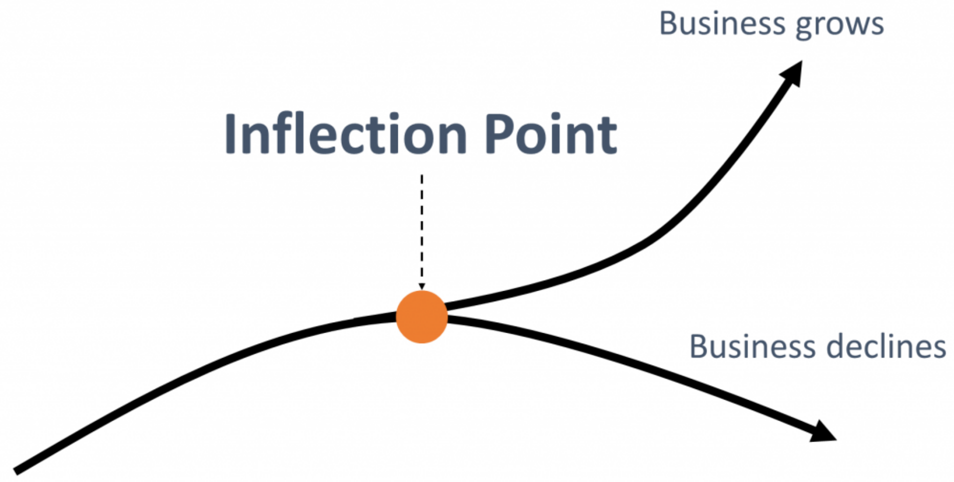
MACHINE PROGRAMMING

ANY TECHNIQUE THAT AUTOMATES SOME ASPECT OF SOFTWARE (AND HARDWARE) DEVELOPMENT.

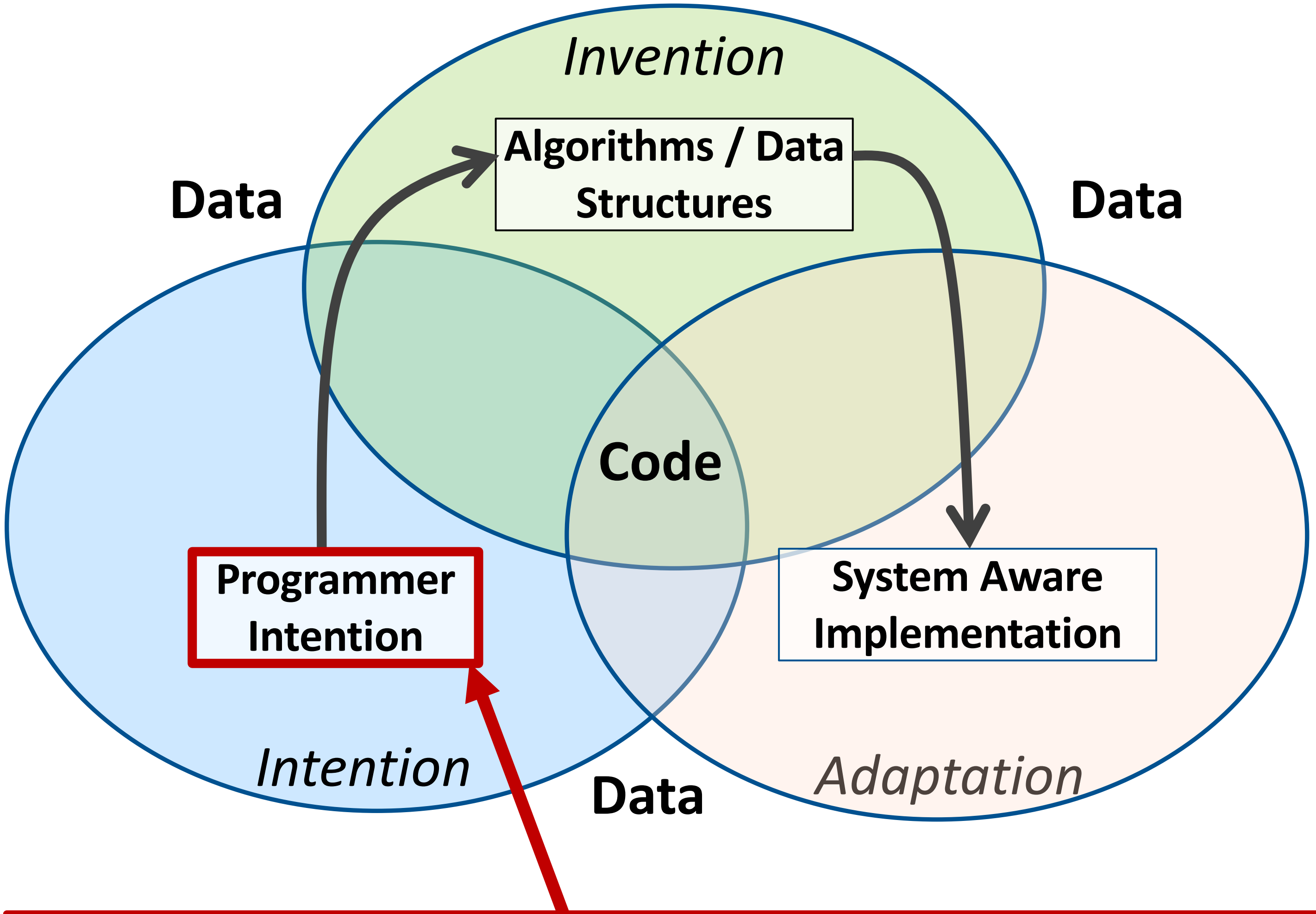
ENABLE THE WORLD TO CREATE SOFTWARE.

The Machine Programming Inflection Point

- **ML/DL algorithms**
  - Transformer, neural recursion, DeepCoder, time series precision and recall, …
- **Parallel hardware**
  - CPU, GPU, NNP, TPU, FPGAs, …
- **Rich & dense programming data**
  - 33,000 github repos in 2008
  - 180,000,000 github repos in 2019
- *And programs are graphs after all*



Three Pillars of Machine Programming *

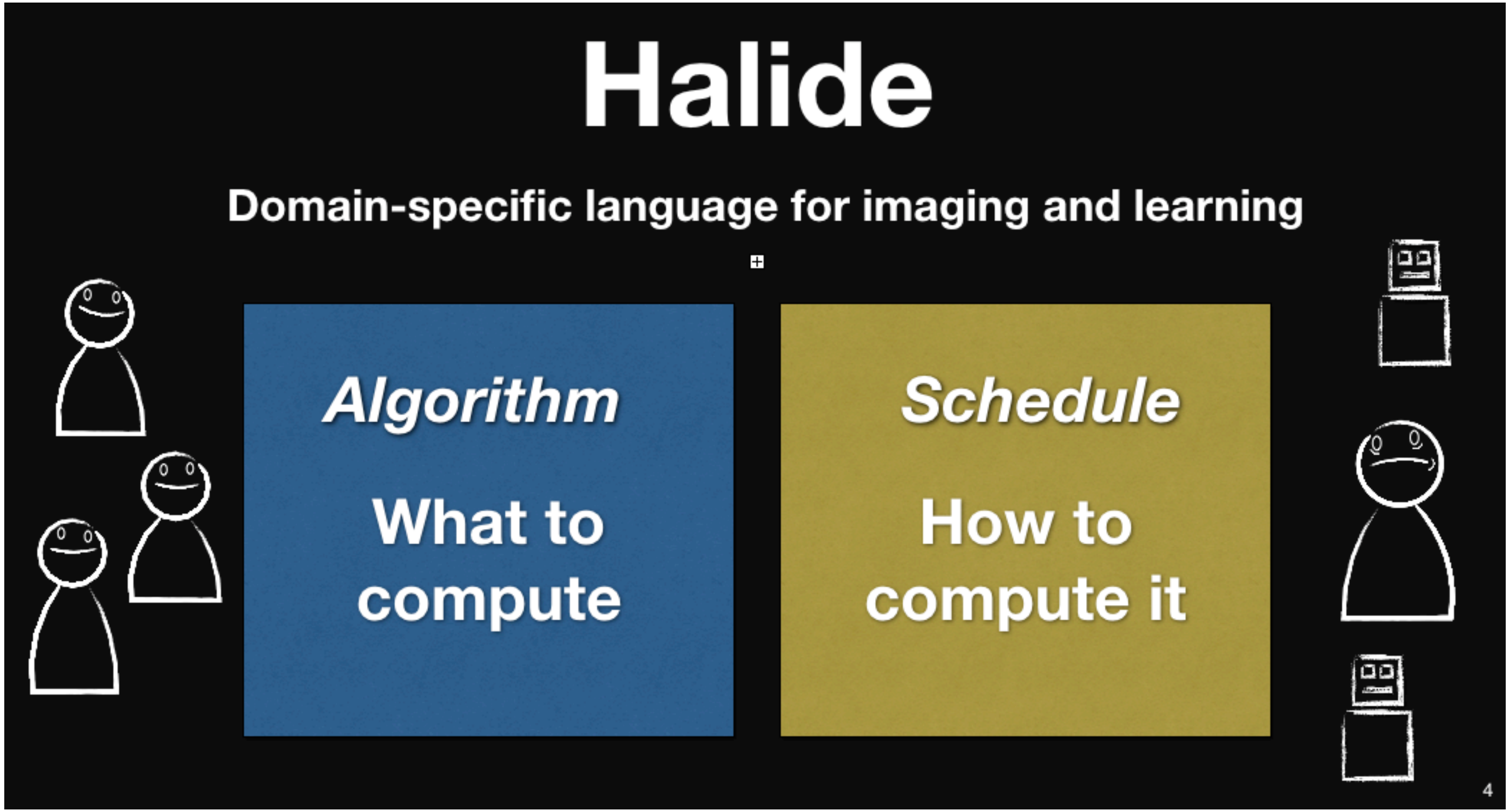


* Justin Gottschlich, Intel
Armando Solar-Lezama, MIT
Nesime Tatbul, Intel
Michael Carbin, MIT
Martin Rinard, MIT
Regina Barzilay, MIT
Saman Amarasinghe, MIT
Joshua B Tenenbaum, MIT
Tim Mattson, Intel

ACM SIGPLAN Workshop on Machine Learning and Programming Languages (MAPL), PLDI'18, arxiv.org/pdf/1803.07244.pdf

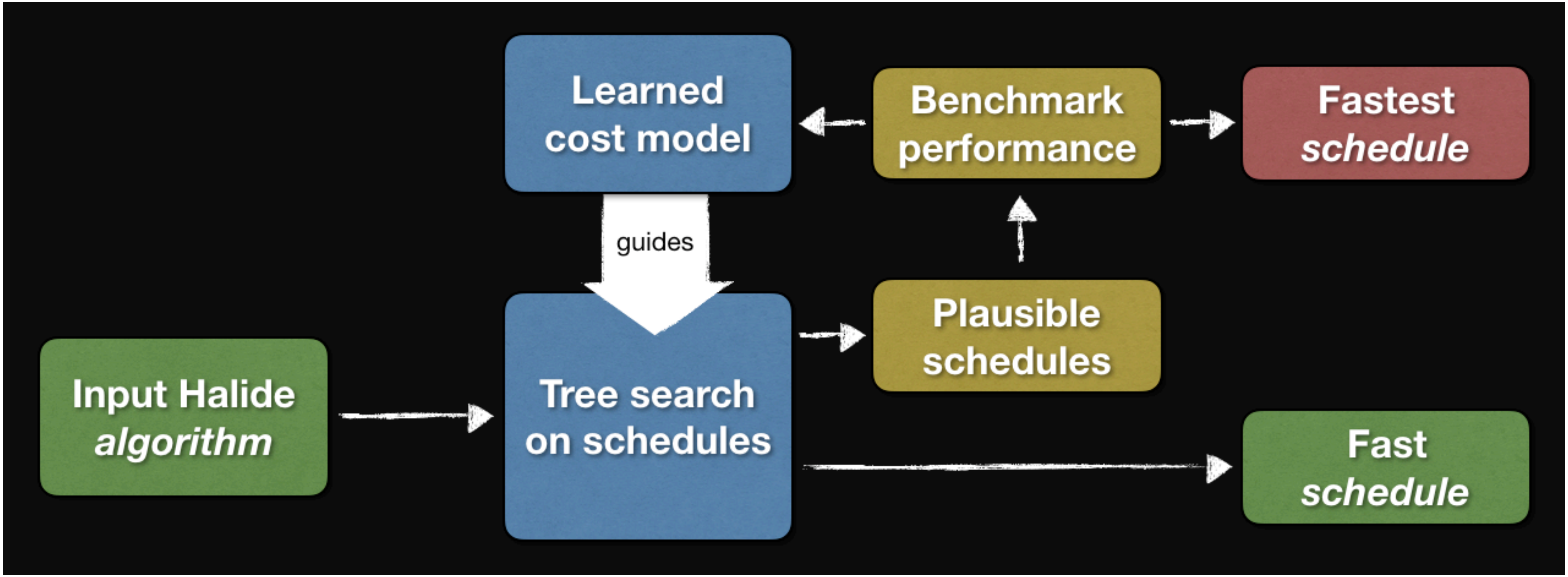
The programmer *only* does this.

Learning to Optimize Halide with Tree Search and Random Programs *

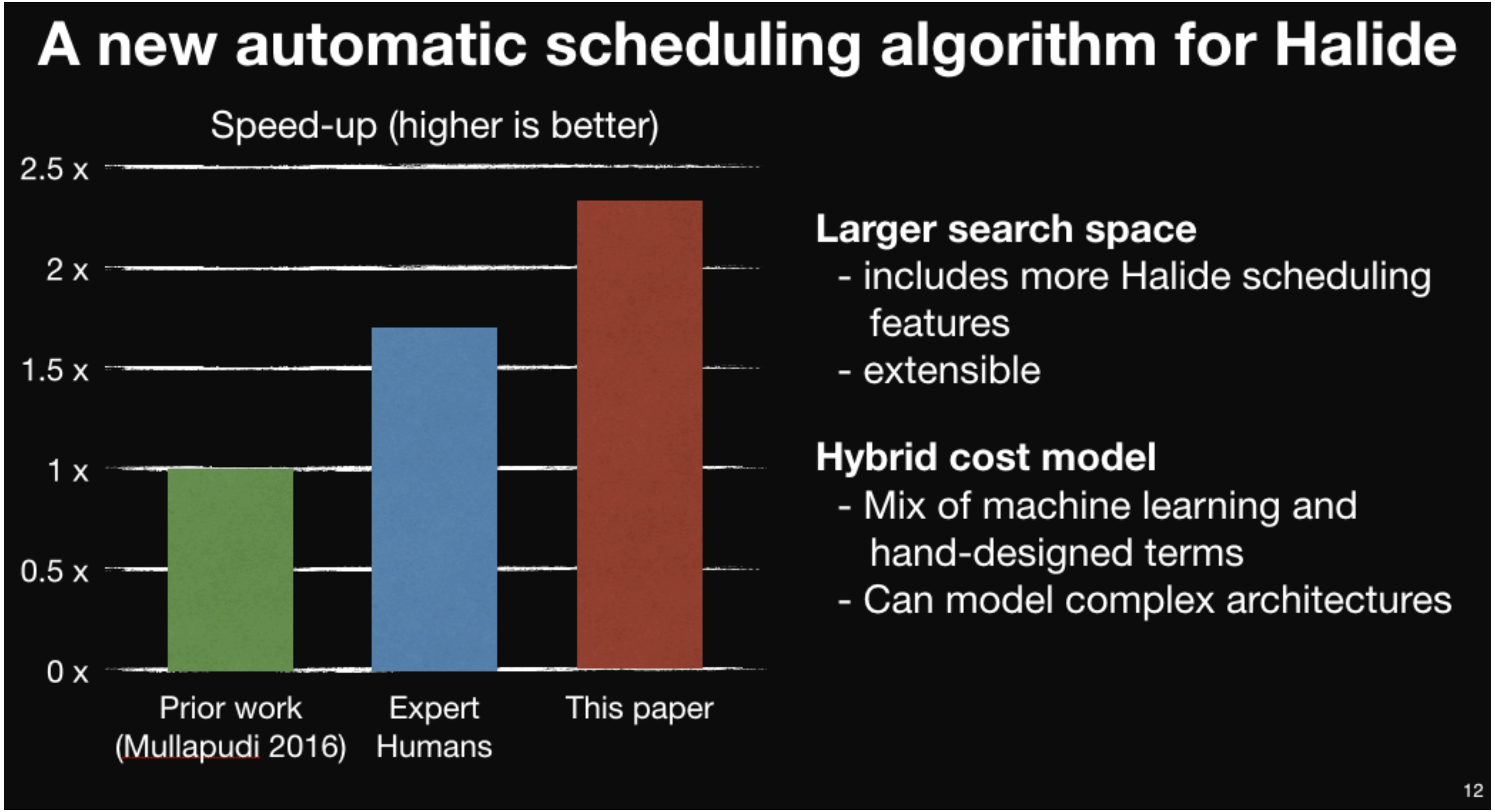


"Learning to Optimize Halide with Tree Search and Random Programs" Andrew Adams, et. al, SIGGRAPH 2019 (https://dl.acm.org/citation.cfm?id=3322967)

Halide Learned System Design



Significantly Outperforms Human Ninjas



“Learning to Optimize Halide with Tree Search and Random Programs” Intel CAPA funded, SIGGRAPH 2019 (https://dl.acm.org/citation.cfm?id=3322967)

*Neo: A Learned end-to-end Query Optimizer **

Traditional Optimizers vs. Neural Optimizer

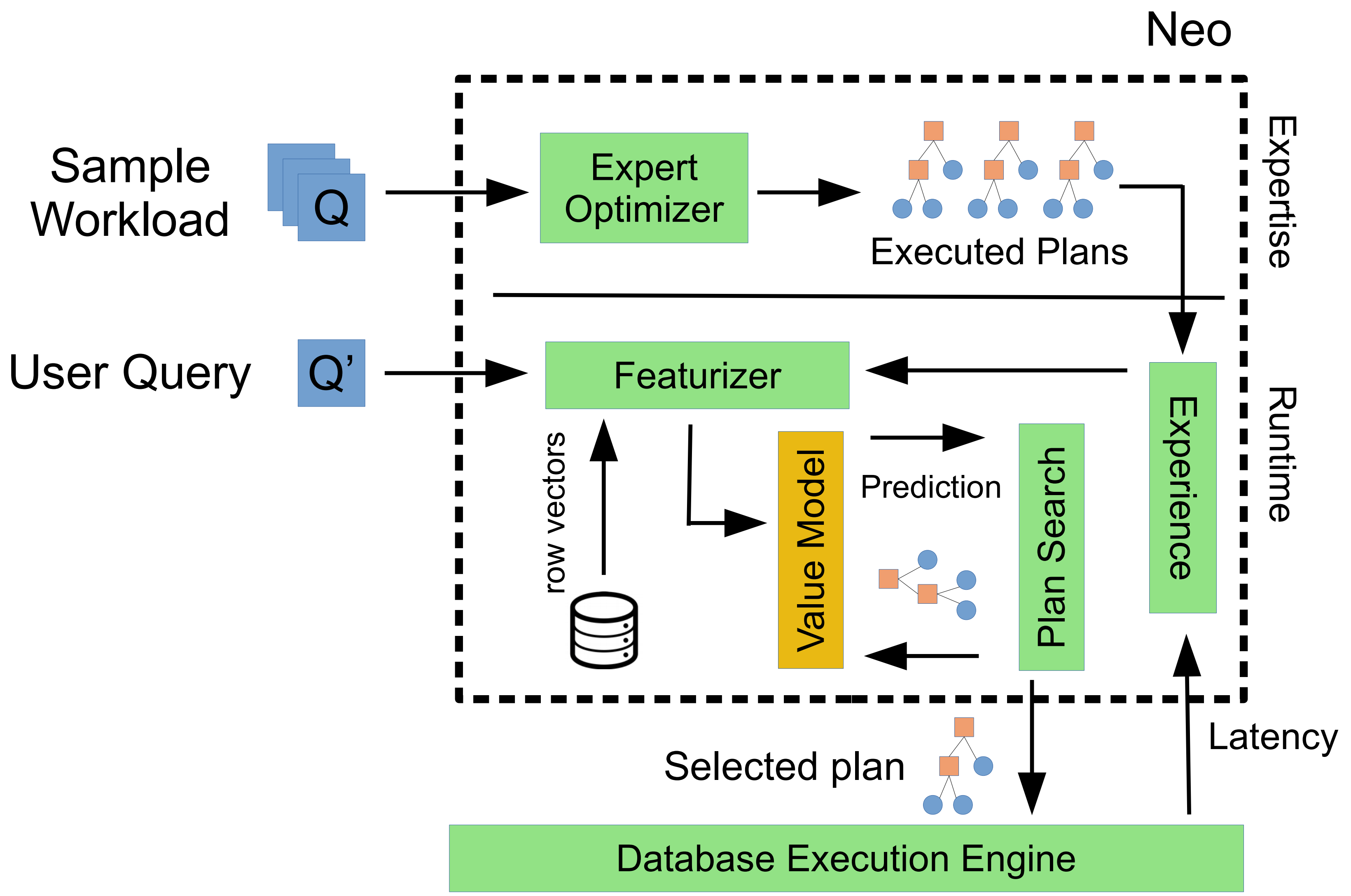
	Traditional Optimizers	Neural Optimizer (Neo)
Creation	Human developers	Learning from demonstration, Reinforcement learning
Query Representation	Operator tree	Tree with feature encodings
Cost Model	Hand-crafted model	Learned DNN model (value network)
Plan Space Search	Enumeration, Heuristics, Dynamic programming	DNN-guided search strategy (best-first)
Cardinality Estimation	Histograms, Hand-crafted models	Histograms, Learned embeddings (row vectors)

Neo: first fully learned end-to-end solution to DB query optimization:

Matches/exceeds state of the art open-source & commercial optimizers within 24 hrs of training

* ***"Neo: A Learned Query Optimizer"*** (Marcus et al., VLDB '19 http://www.vldb.org/pvldb/vol12/p1705-marcus.pdf

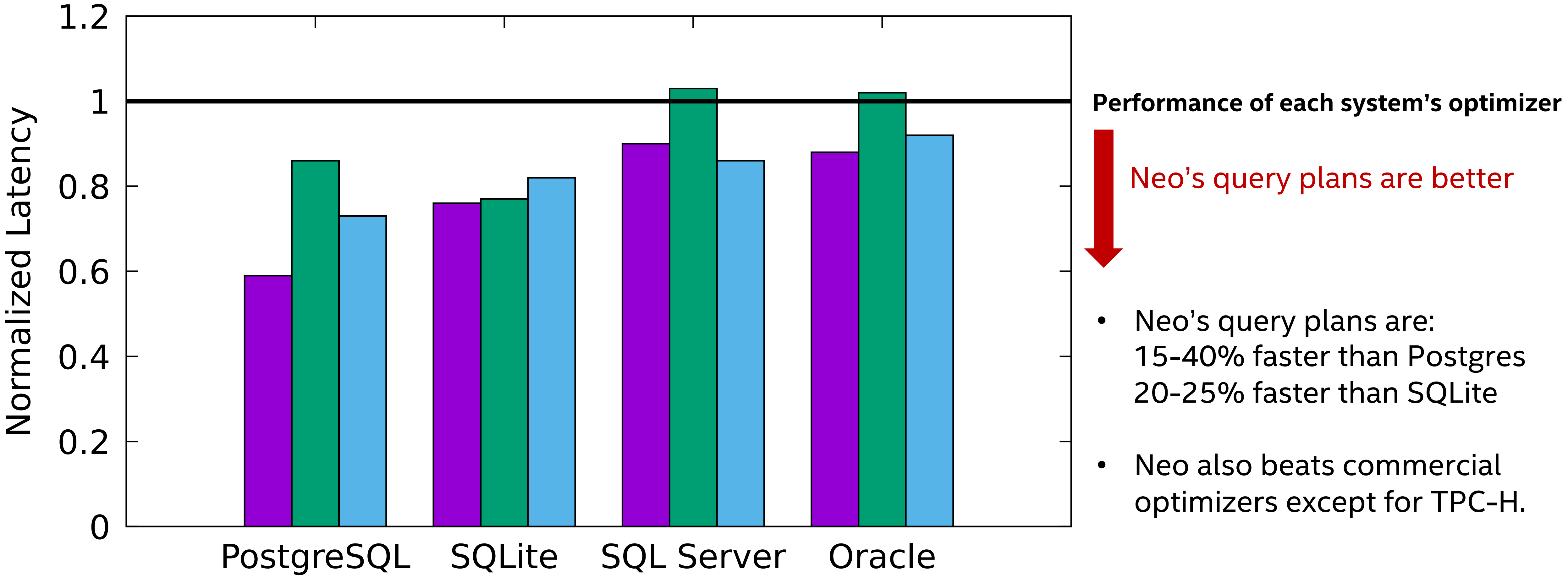
Neo's Learning Framework



Key ingredients:

- Learning from demonstration
- Value Model: DNN based on tree convolution
- Reinforcement learning
- Query and plan level feature encoding techniques (e.g., row-vectors that capture predicate semantics)

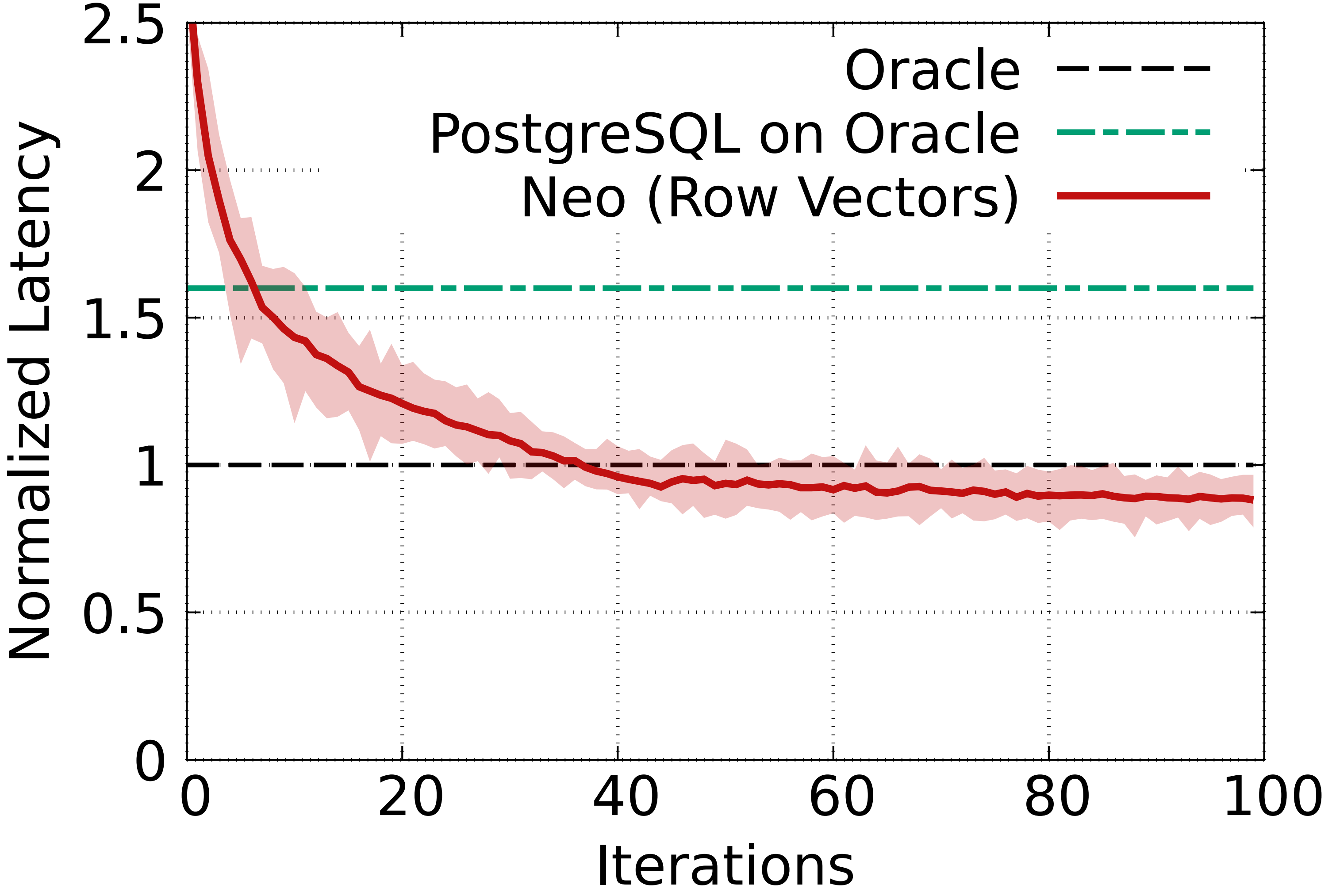
Neo vs. Traditional Optimizers: Overall Performance



JOB — Join Order Benchmark over the IMDB database, complex predicates, lots of column correlations
TPC-H — Standard OLAP benchmark from TPC at scale factor 10
Corporation — 2 TB dataset with 8000 unique queries from a large DB corporation

- Neo's query plans are:
  15-40% faster than Postgres
  20-25% faster than SQLite
- Neo also beats commercial optimizers except for TPC-H.

Neo vs. Oracle Optimizer: Learning Curve



CSAIL

MIT Computer Science & Artificial Intelligence Lab

Internal | Login

Research People News Events Admissions Engage About

‹ BACK TO NEWS

October 03 '18

Google, Intel and Microsoft team up w/CSAIL on new data-driven initiative

WRITTEN BY

Adam Conner-Simons

RELATED

DSAIL homepage

CSAIL Alliances program

Tim Kraska

DSAIL Launched Jan 2019 @ MIT in collaboration with Google-Microsoft <link>

Summary

We are at an unprecedented convergence of massive compute with massive data …
This confluence will have a lasting impact on both how we do computing and what computing can do for us!

Questions?

Notice and Disclaimers

Notice: This document contains information on products in the design phase of development. The information here is subject to change without notice. Do not finalize a design with this information. Contact your local Intel sales office or your distributor to obtain the latest specification before placing your product order.

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL® PRODUCTS. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER, AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY RELATING TO SALE AND/OR USE OF INTEL PRODUCTS, INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT, OR OTHER INTELLECTUAL PROPERTY RIGHT. Intel products are not intended for use in medical, life saving, or life sustaining applications. Intel may make changes to specifications, product descriptions, and plans at any time, without notice.

All products, dates, and figures are preliminary for planning purposes and are subject to change without notice.

Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined.“ Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them.

Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance.

The Intel products discussed herein 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.

Copies of documents which have an order number and are referenced in this document, or other Intel literature, may be obtained by calling 1-800-548-4725, or by visiting Intel's website at http://www.intel.com.

Notice and Disclaimers Continued …

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

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice. Notice revision #20110804