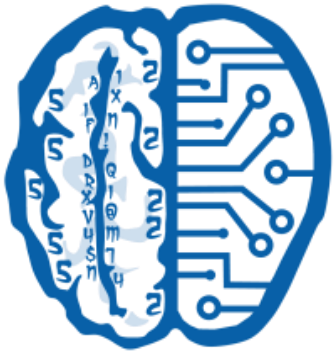


ICRI-CI

Intel Collaborative Research Institute
Computational Intelligence

intel

Capstone Overview

Architecture for Big Data & Machine Learning

Debbie Marr

ICRI-CI 2015 Retreat, May 5, 2015

INTEL LABS

Deliver breakthrough innovations to fuel Intel's growth and technology leadership

Accelerators

Memory Traffic Reduction

Memory Intensive Arch.

Context-based Prefetching

Deep Learning

SimNets

Distributed Methods for Deep Learning

Scene Understanding

Saliency Estimation

Statistics of Depth Images

Arguments for Persuasive Discussion

Universal Semantics

Transcript Quality

Inference for NLP

Relations and Events

Extraction Knowledge Graphs

Hybrid Models

Syntactic & Semantic Reranking

Language Modeling

2nd-order Embedding

Mental Phenotyping

Reinforcement Learning

ICRI-CI

Intel Collaborative Research Institute Computational Intelligence

Distributed Deep Learning Library

Conversational Speech Understanding

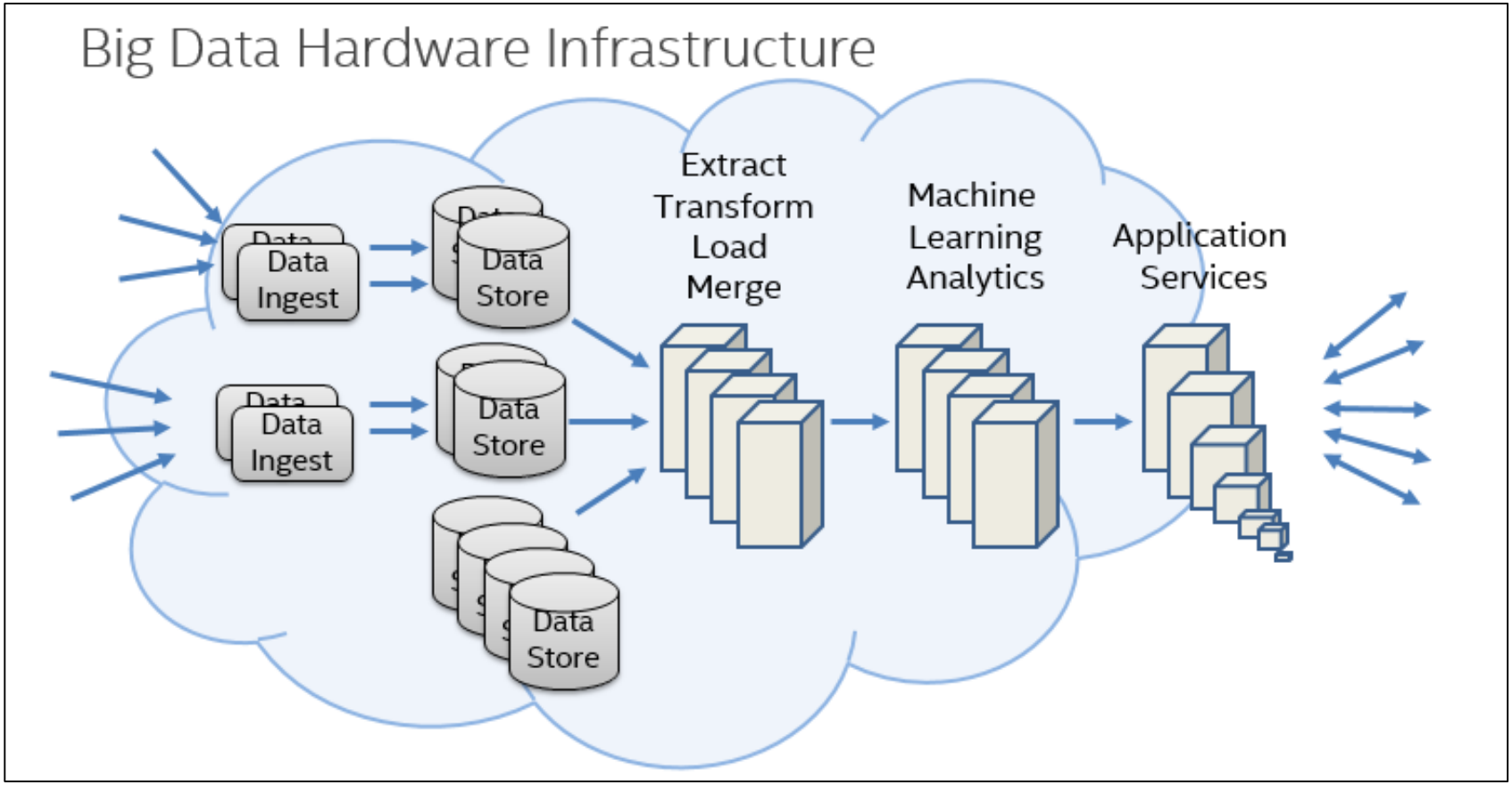
ICRI-CI
Intel Collaborative Research Institute
Computational Intelligence

Accelerators for Big Data and Machine Learning

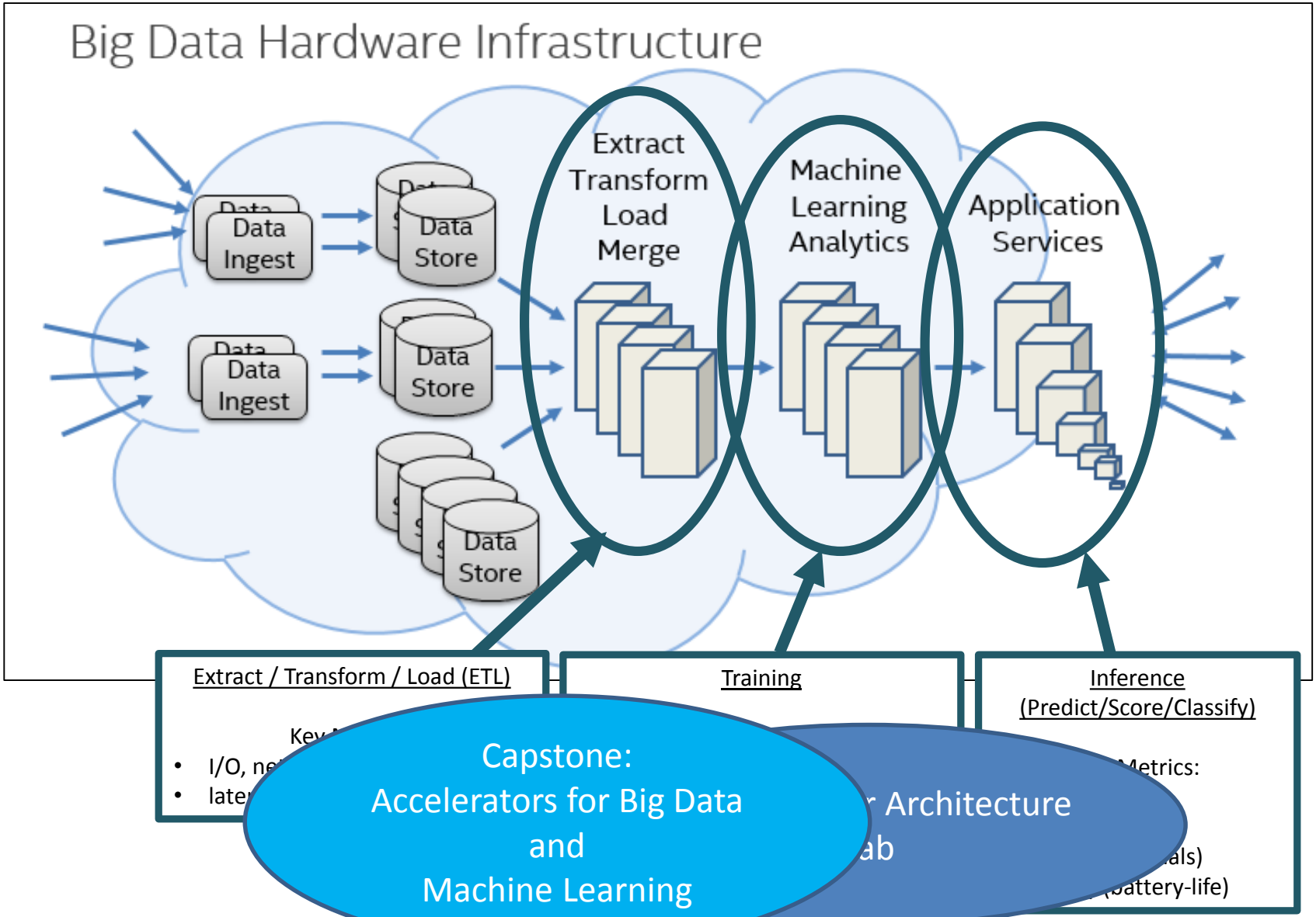
INTEL LABS

Deliver breakthrough innovations to fuel Intel's growth and technology leadership

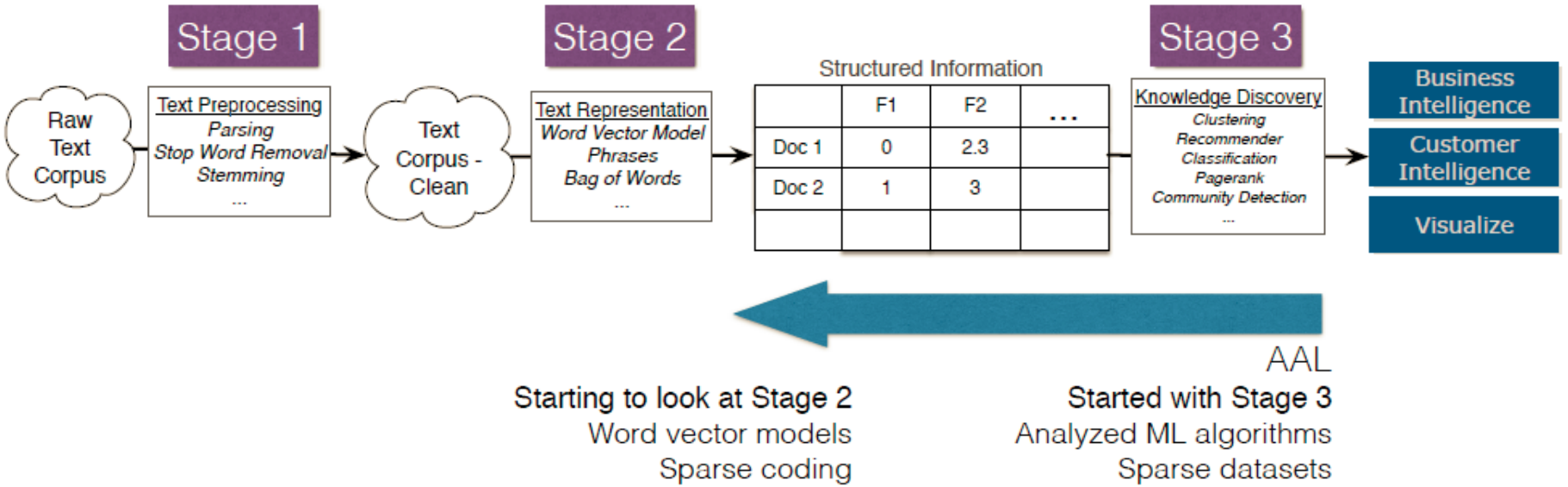
Big Data / Machine Learning Hardware Infrastructure View



AAL & ICRI-CI Accelerator Investments



Example: Text Analytics Pipeline



AAL

Starting to look at Stage 2
Word vector models
Sparse coding

Started with Stage 3
Analyzed ML algorithms
Sparse datasets

ICRI

Start with Stage 1
Analysis on LARGE amounts of raw text
Map Algorithm —> Data storage
(Disk/SSD/DRAM/Cache)

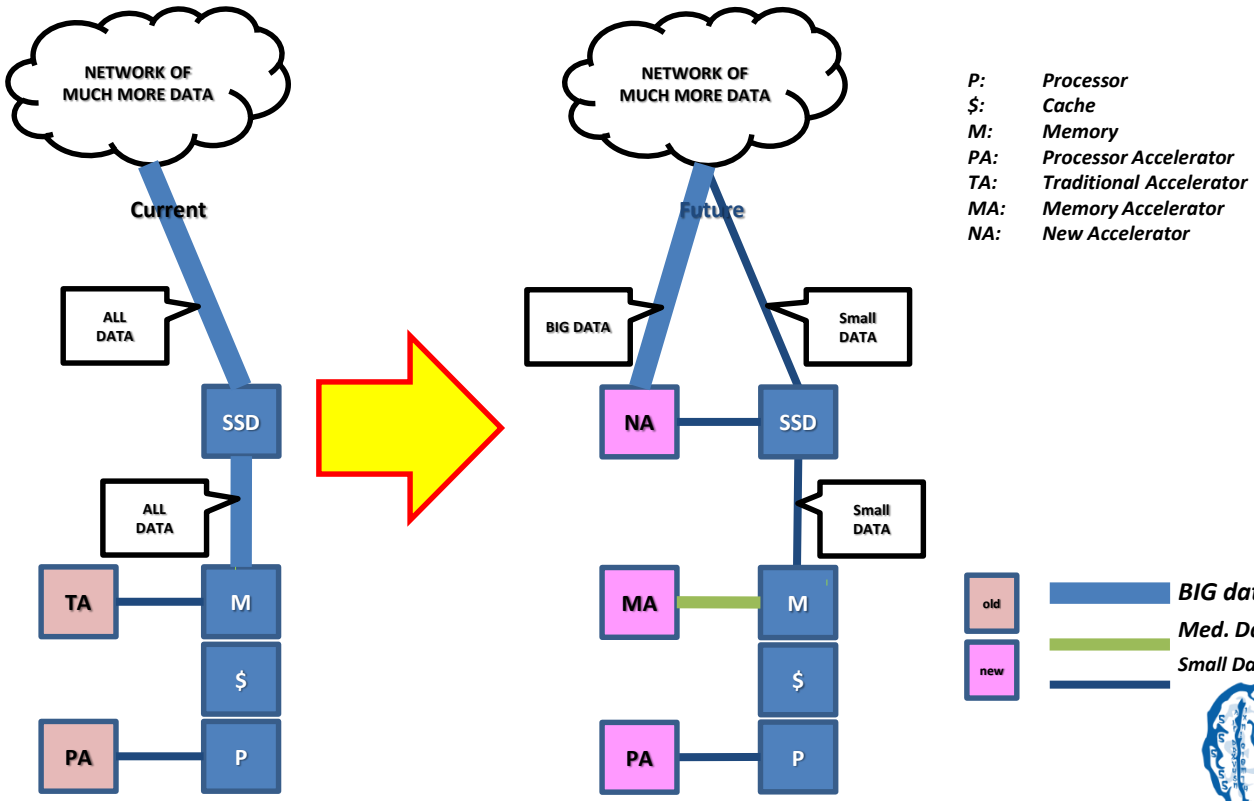
ICRI-CI
Intel Collaborative Research Institute
Computational Intelligence
intel

Capstone

Optimized IA for Big Data & Machine Learning

Goal: Break-through performance and energy-efficiency for a big data analytics platform

1. Data movement within/across nodes, where and when to (not) store
2. Computation placed in the storage & network hierarchy
3. New accelerators for big data
4. Applications and usage of new memory technologies (e.g. memristors)
5. Leveraging ML algorithms for new microarchitectures



ICRI-CI
Intel Collaborative Research Institute
Computational Intelligence
intel

Plan and Timeline

Goal: Break-through performance and energy-efficiency for a big data analytics platform

Time	Plan	Status
Q1'15	Education, background, select target & workloads.	Bi-weekly meetings since Q4'14: * Education & background: target area and algorithms * Looked at Hi-Bench and search for ETL workloads * High-level analysis of workloads.
Q2'15	Broad-stroke microarchitecture, performance, and workloads	2 F2F meetings setup: * 5/3 at the Technion * 6/11 at Intel Jones Farm
Q3'15	Next-level of detail for microarchitecture, performance, and workloads	
Q4'15	High-level simulations and models of workloads on microarchitecture	
Q1'16	Detailed simulations and models of workloads on microarchitecture	
Q2'16	Bring simulator and workloads in-house for further analysis and in-house assessment	
Q3'16 -Q2'17	Parallel work on accelerators for target workloads and microarchitecture.	

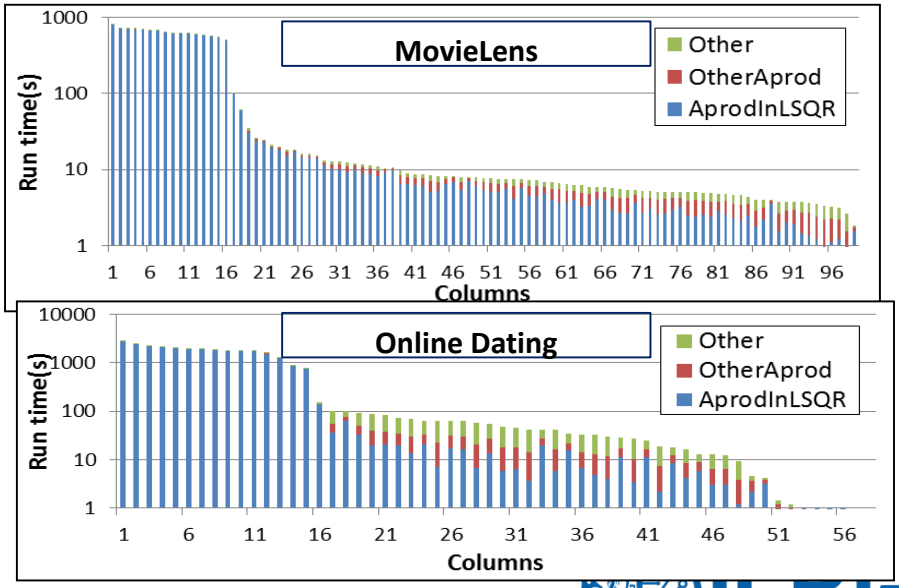
ICRI-CI
Intel Collaborative Research Institute
Computational Intelligence
intel

Example: Algorithms -> Accelerators

Step 1: Study algorithm

- Study algorithm, usage, alternatives, benefits, trade-offs
- Get multiple code & datasets
- Analyze code on several datasets

$$\begin{aligned} \underset{\mathbf{w}_j}{\text{minimize}} \quad & \frac{1}{2}\|\mathbf{a}_j - A\mathbf{w}_j\|_2^2 + \frac{\beta}{2}\|\mathbf{w}_j\|_2^2 + \lambda\|\mathbf{w}_j\|_1 \\ \text{subject to} \quad & \mathbf{w}_j \geq 0 \\ & w_{j,j} = 0, \end{aligned}$$

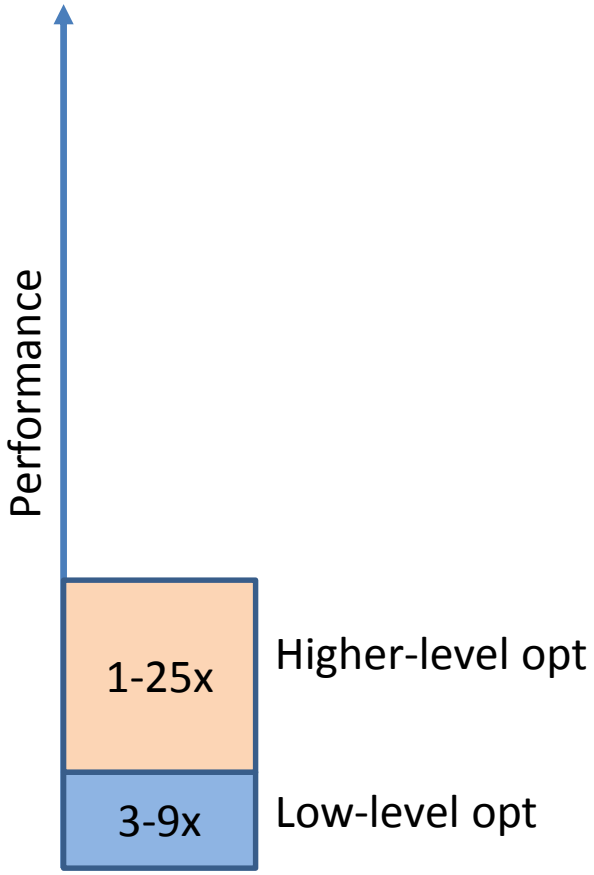


Example: Algorithms -> Accelerators

Step 2: Optimize algorithm

- Low-level optimizations
  - Vectorizing
  - Threading (better)
  - Software prefetches
  - Branch prediction
- Higher-level optimizations
  - Algorithm & Data format changes
  - Compression
  - Re-ordering/pipelining phases

→ Identify bottlenecks & pain points

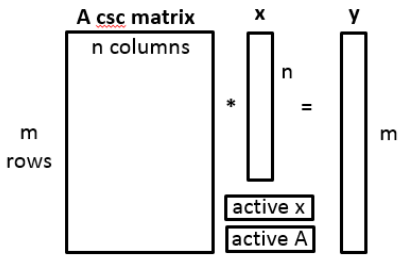


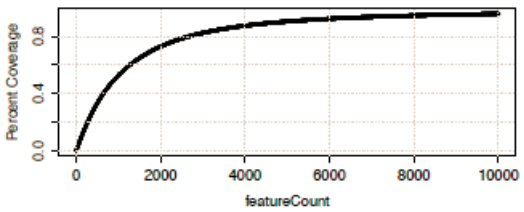
ICRI-CI
Intel Collaborative Research Institute
Computational Intelligence
intel

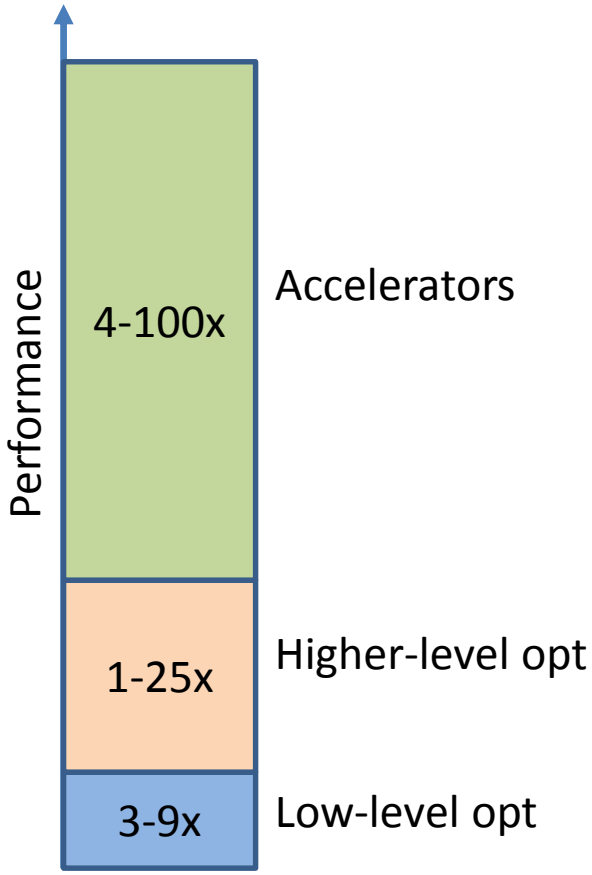
Example: Algorithms -> Accelerators

Step 3: Accelerators

- Compute characteristics
  - Serial
  - Parallel
  - Operations
- Bandwidth characteristics
  - Caching vs. streaming
  - Gathers/scatters
  - Working set size
- Communication





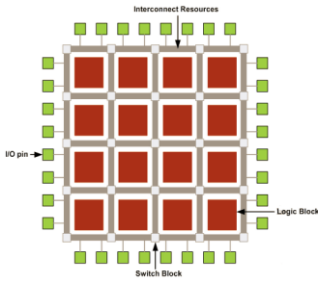


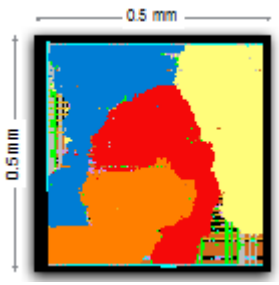
ICRI-CI
Intel Collaborative Research Institute
Computational Intelligence
intel

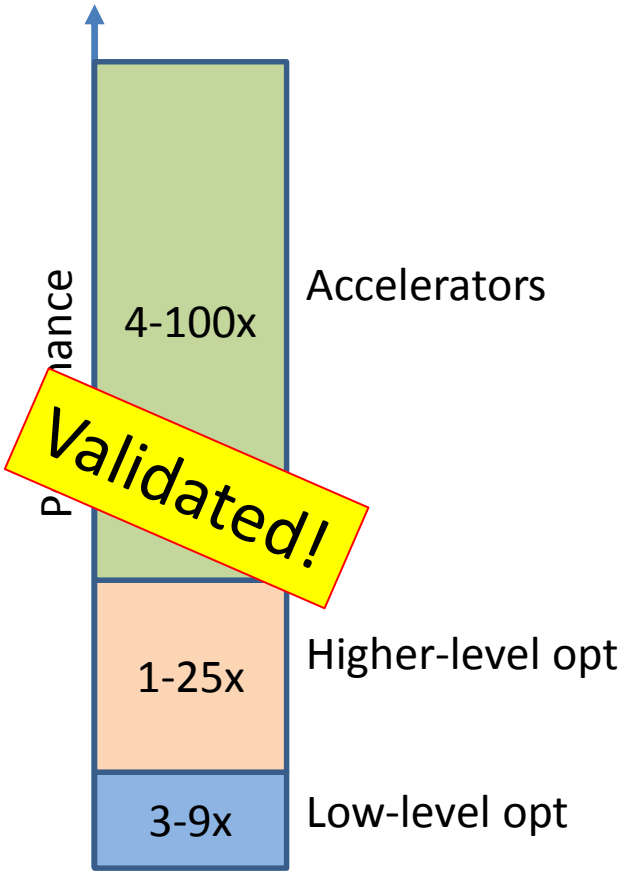
Example: Algorithms -> Accelerators

Step 4: Prototype

- FPGA
  - Modify program to offload work to FPGA
  - Accelerator functionality
- Synthesis
  - Power, Area, Frequency
- Simulations
  - Put it all together inc. overhead
- Programmability
  - New instruction?
  - PCIe device?
- Competitive Analysis
- Transfer to product

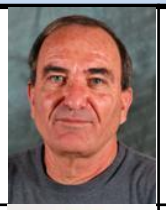






ICRI-CI
Intel Collaborative Research Institute
Computational Intelligence
intel

ICRI-CI Architecture Track – Year 4-5

Reduction of Memory traffic for Big Data			
Funnel:	*Reduction of Data Movement in Big Data system*		**Prof. Uri Weiser** Prof. Avinoam Kolodny
Accelerators for Big Data & Machine Learning	*Novel Accelerators*		**Prof. Ran Ginosar** Prof. Oded Schwartz
Machine Learning for Architecture	*Adaptive Systems*		**Prof. Yoav Etsion** Prof. Uri Weiser Prof. Shie Mannor
Memory Intensive Architecture: **New memory based machine**	*New Technology based Architecture*		Dr. Shahar Kvatinsky **Prof. Avinoam Kolodny** Prof. Eby Friedman (Technion/Rochester) Prof Yuval Cassuto

ICRI-CI
Intel Collaborative Research Institute
Computational Intelligence