



Capstone Overview Architecture for Big Data & Machine Learning

Debbie Marr

ICRI-CI 2016 Retreat, May 24, 2016

INTEL LABS

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



Accelerators

Universal Semantics

Memory Traffic Reduction

Transcript Quality

Memory Intensive Arch.

Context-based Prefetching

Inference for NLP

Deep Learning

Relations and Events

SimNets

Extraction Knowledge Graphs

Distributed Methods for Deep Learning

Hybrid Models

Scene Understanding

Syntactic & Semantic Reranking

Saliency Estimation

Language Modeling

2nd-order Embedding

Statistics of Depth Images

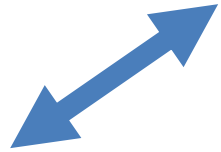
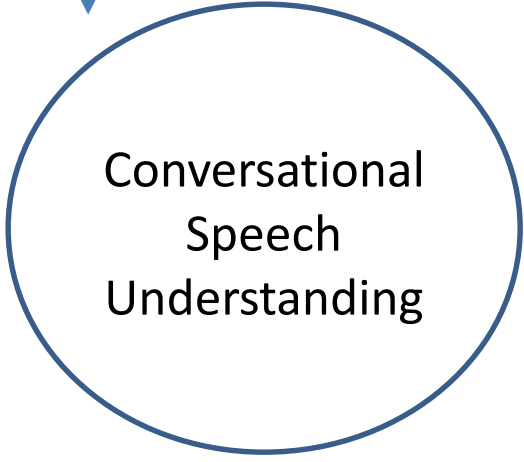
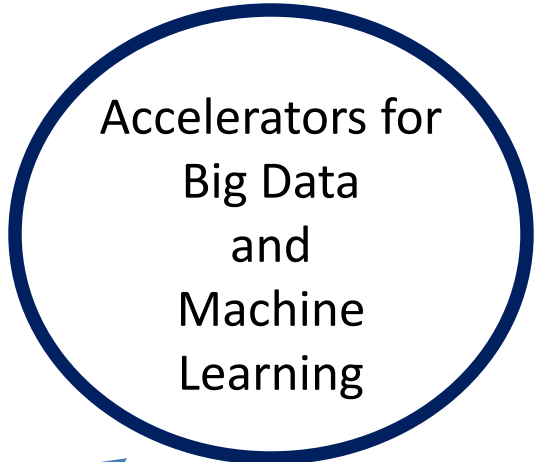
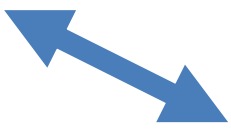
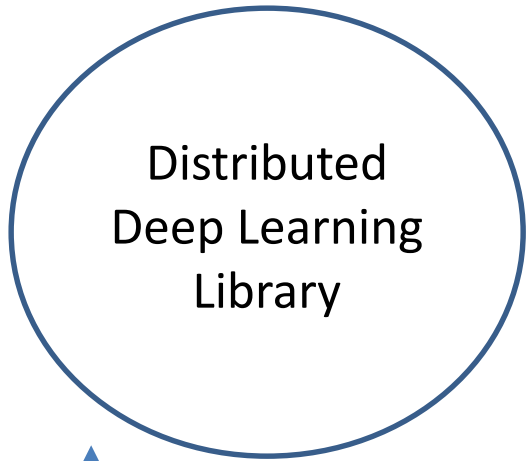
Mental Phenotyping

Arguments for Persuasive Discussion

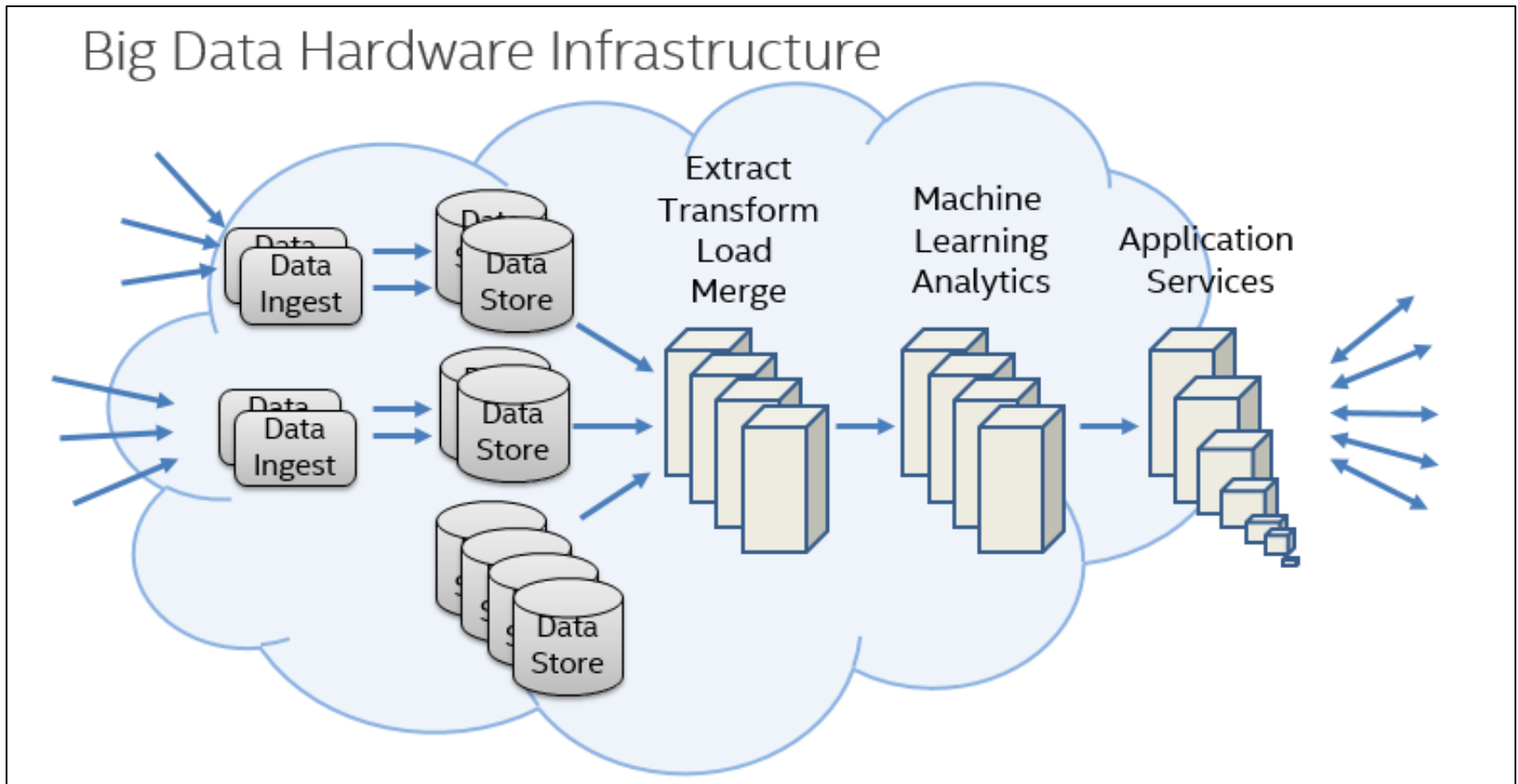
Reinforcement Learning



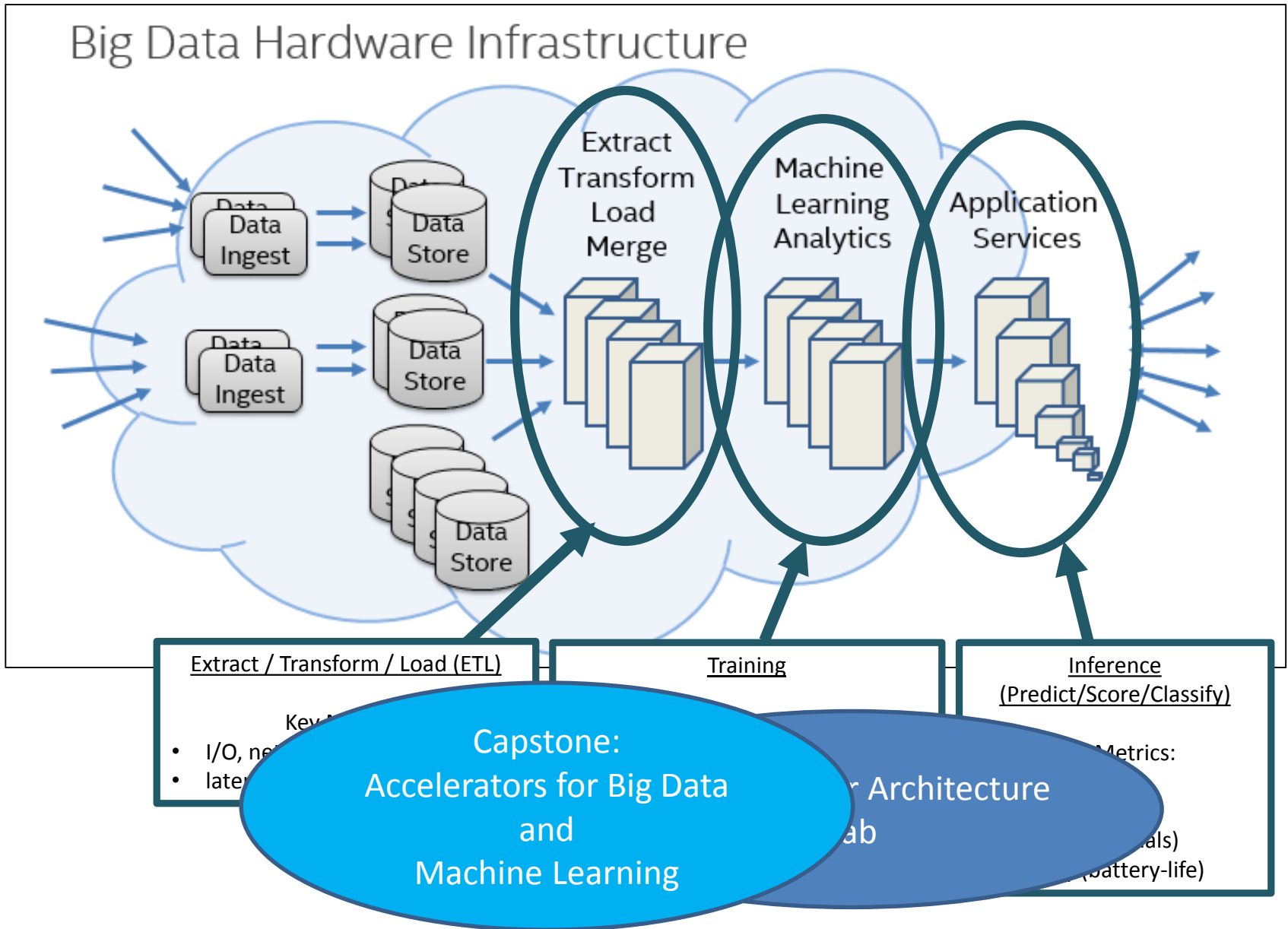
INTEL LABS
Deliver breakthrough innovations to fuel Intel's growth and technology leadership



Big Data / Machine Learning Hardware Infrastructure View

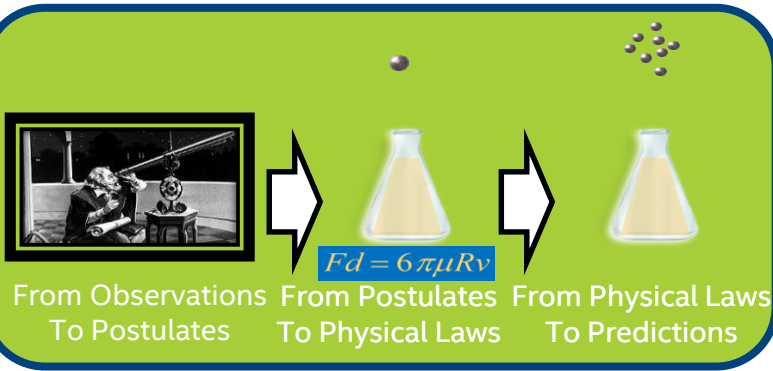


AAL & ICRI-CI Accelerator Investments

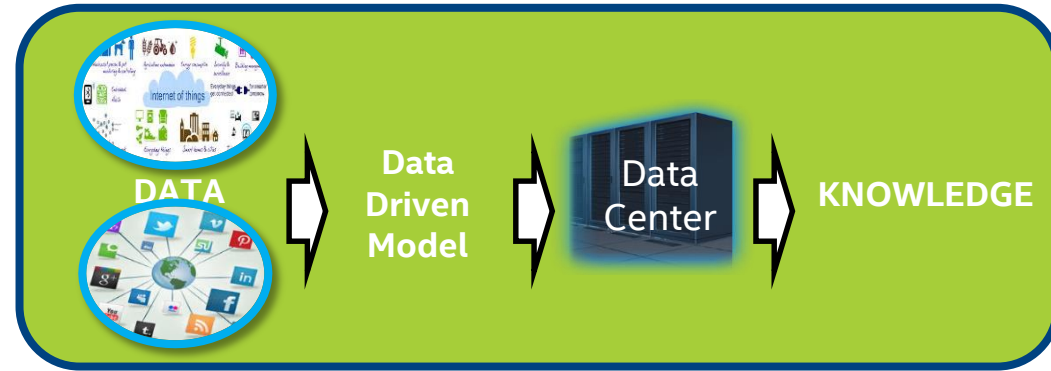


Humans vs. Compute Systems

Human Knowledge Acquisition



Compute Systems Knowledge Acquisition



Brain:

Compute efficiency: 4-5 Gop/W

Power of compute vs. communication: 50:50

Computing systems:

Compute efficiency: 4-5 Gop/W achievable

Power of compute vs. communication: 10:90

Refactor
compute
systems around
data &
communication

- This year data center flops: ~ 1 Exaflop (10^{18})
- This year internet data rate: ~ 1.6 Zetabyte/s (10^{14})
 - $\sim 10^4$ flops to process 1 byte of internet traffic
- Compute doubling every 1.5 years
- Data storage doubling every 1 year
- Data analytics fastest growing class of data center workload

Computing Systems:

Don't work harder, work smarter

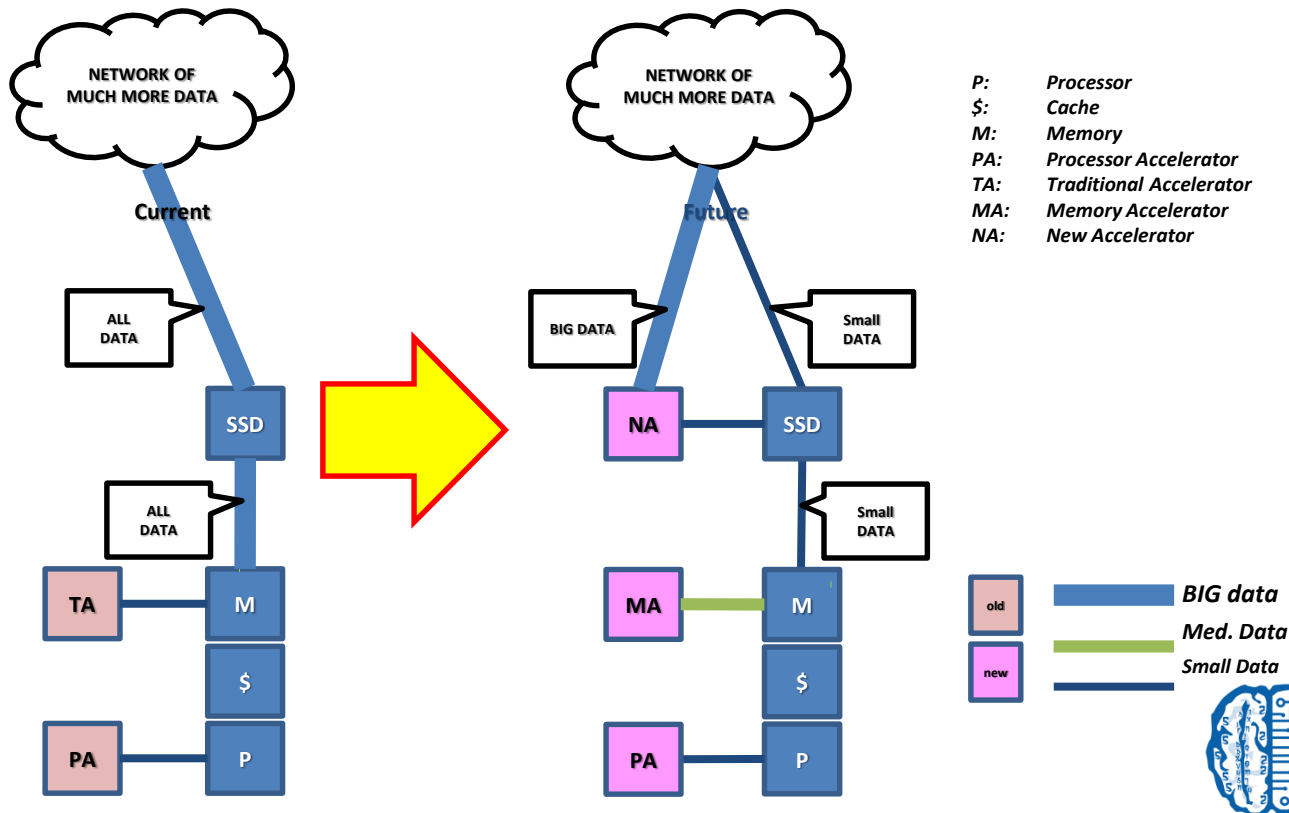
- Working harder:
 - Peak FLOPs
 - Peak Bandwidth
- Working smarter:
 - Re-think compute, storage, bandwidth
 - Move data to compute, or compute to data?
 - Dynamically adapt to changing needs for data, compute, bandwidth
 - Leverage the sparsity in the models to our advantage.

Capstone

Optimized IA for Big Data & Machine Learning

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

1. Data movement in/across nodes
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



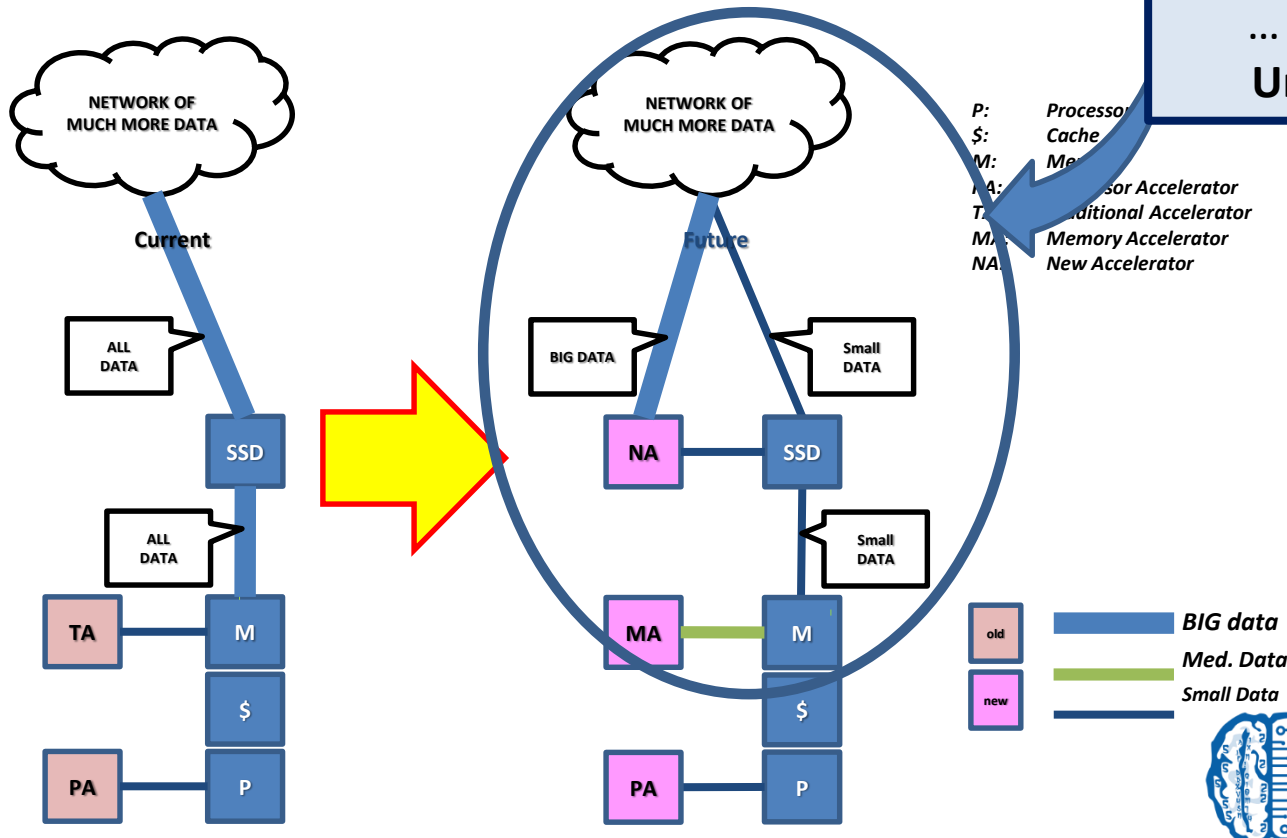
Capstone

Optimized IA for Big Data & Machine Learning

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

1. Data movement in/across nodes
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

“Process-in-storage ... or not?”
Uri Weiser

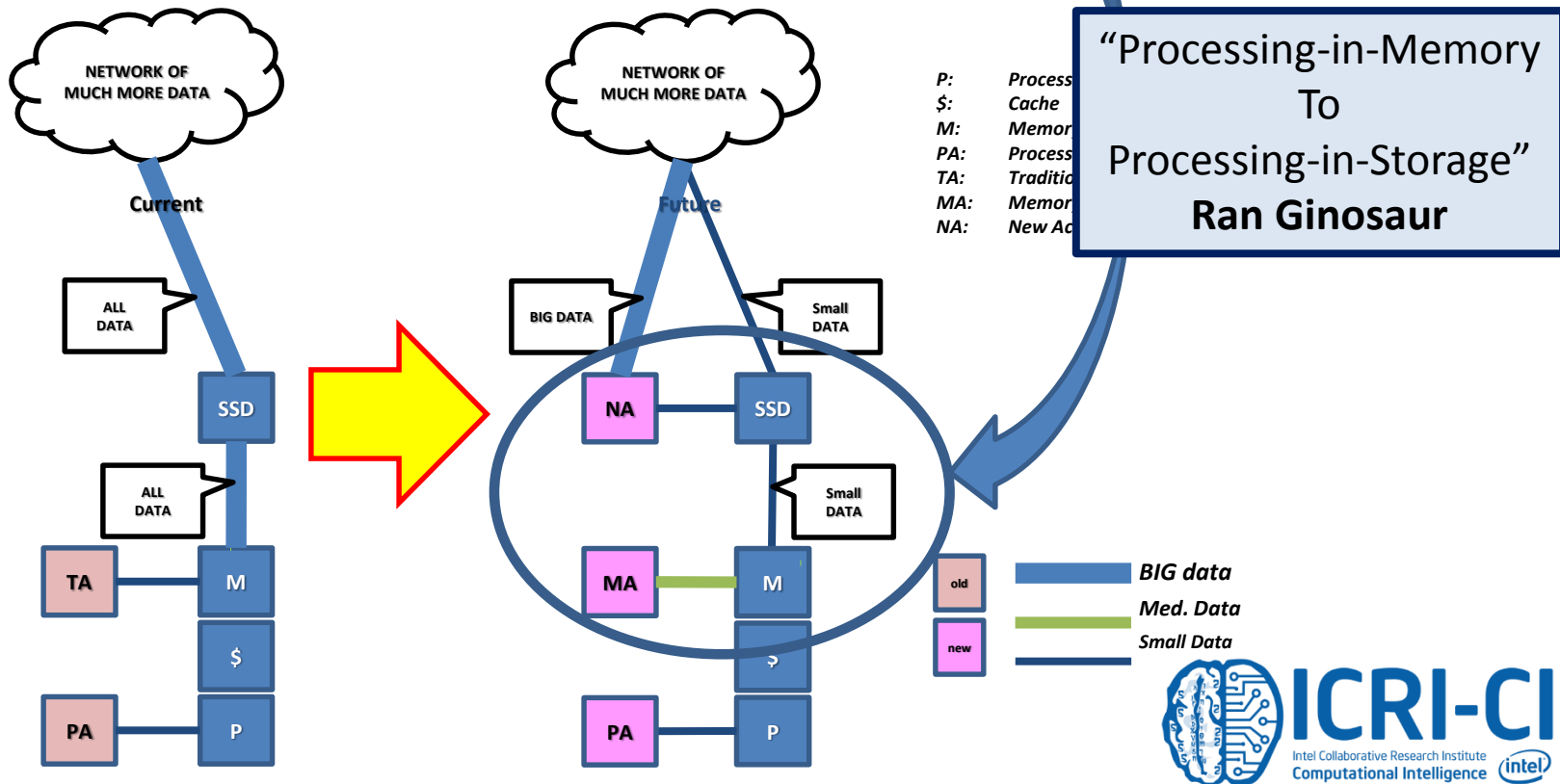


Capstone

Optimized IA for Big Data & Machine Learning

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

1. Data movement in/across nodes
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

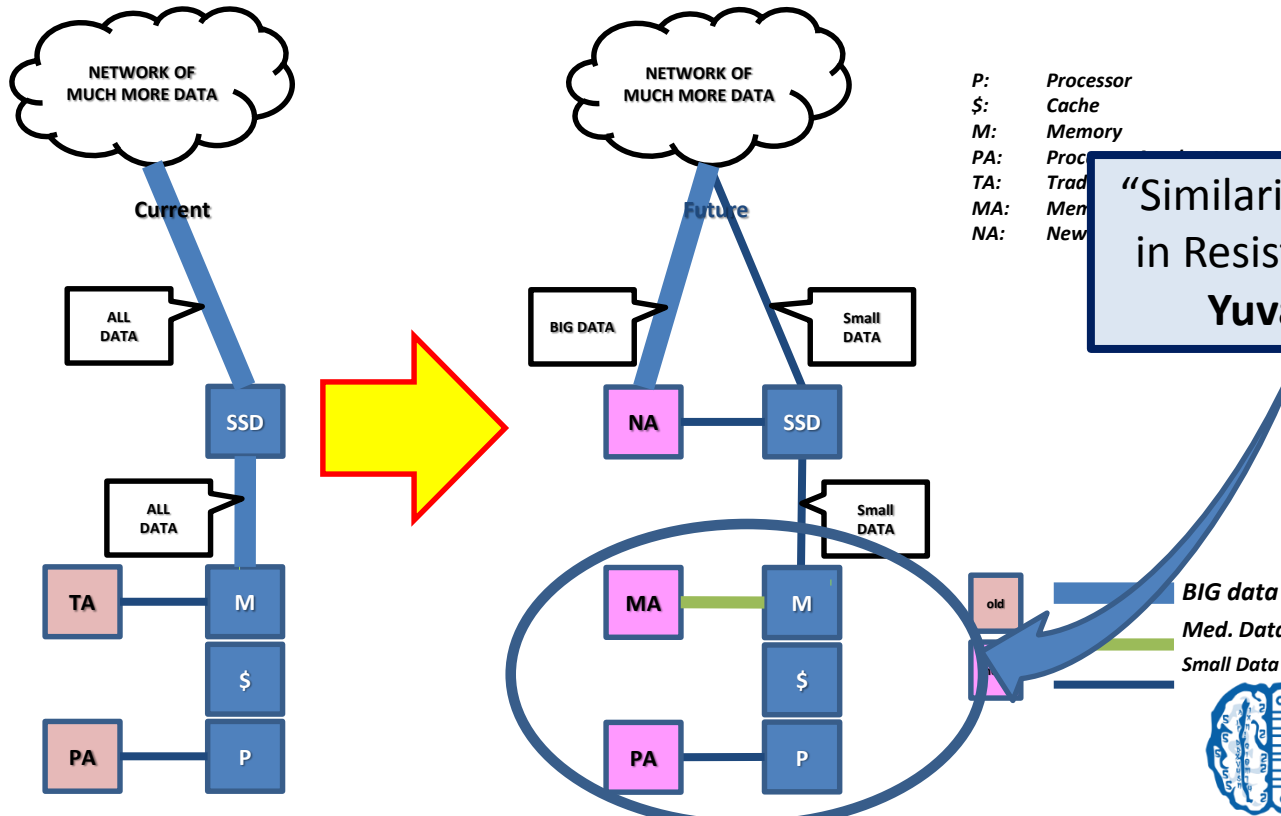


Capstone

Optimized IA for Big Data & Machine Learning

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

1. Data movement in/across nodes
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



“Similarity Calculations in Resistive Memory”
Yuval Cassuto

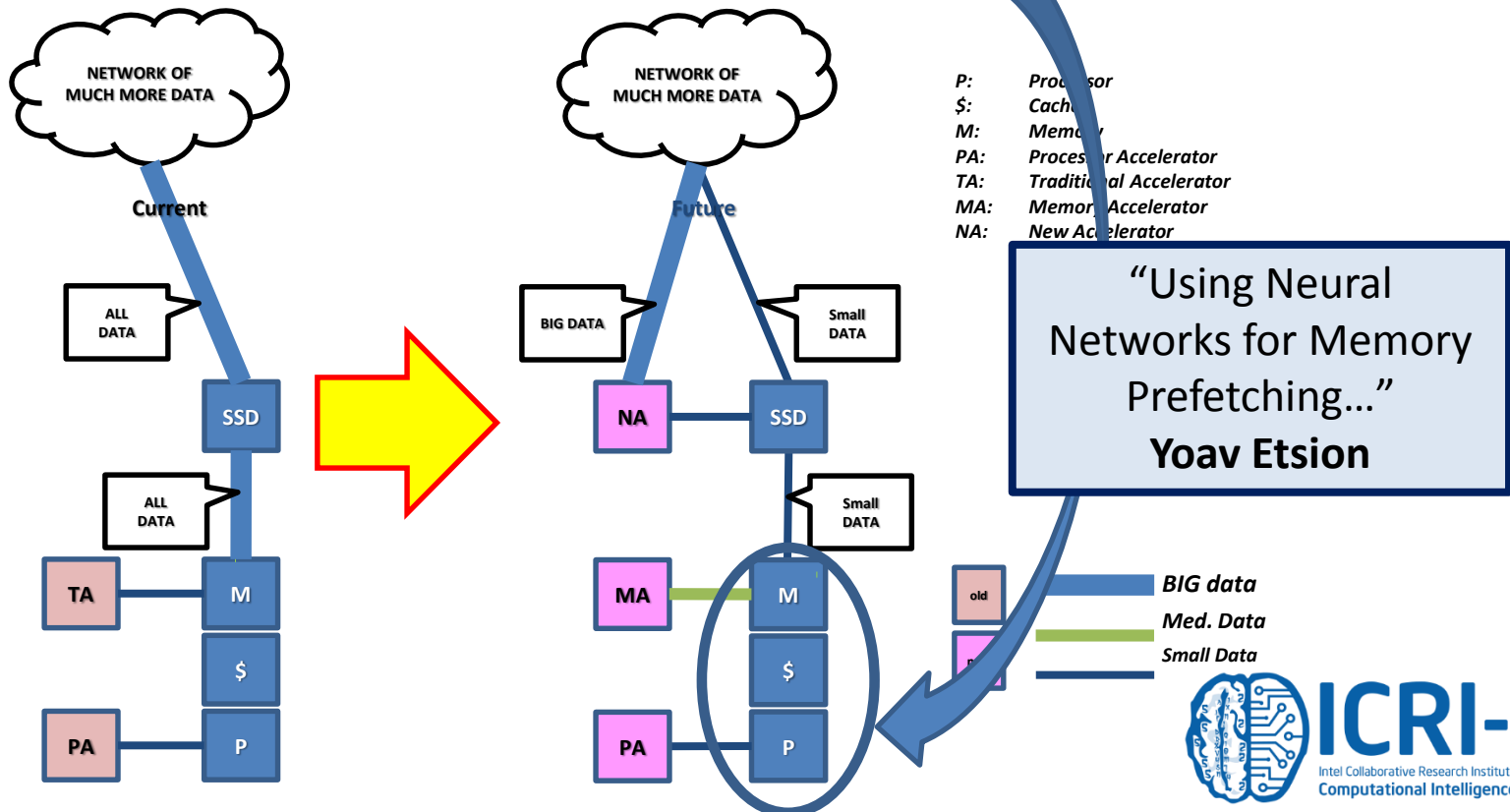
Capstone

Optimized IA for Big Data & Machine Learning

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

1. Data movement in/across nodes
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 Architecture Track – Year 4-5

**Reduction of Memory traffic
and solve Bandwidth System's
bottleneck for Big Data**

Funnel:

*Identify System's
Bandwidth issues
in Big Data
environment and
suggest a remedy*



Prof. Uri Weiser

**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*







Prof. Shahar Kvatinsky
Prof. Avinoam Kolodny
Prof. Eby Friedman (Technion/Rochester)
Prof Yuval Cassuto

BACKUPS

ICRI-CI Architecture Track – Year 4-5

Update this foil

Reduction of Memory traffic and solve Bandwidth System's bottleneck for Big Data

<p><u>Funnel:</u></p>	<p><i>Identify System's Bandwidth issues in Big Data environment and suggest a remedy</i></p>		<p>Prof. Uri Weiser</p>
<p><u>Accelerators for Big Data & Machine Learning</u></p>	<p><i>Novel Accelerators</i></p>		<p>Prof. Ran Ginosar Prof. Oded Schwartz</p>
<p><u>Machine Learning for Architecture</u></p>	<p><i>Adaptive Systems</i></p>		<p>Prof. Yoav Etsion Prof. Uri Weiser Prof. Shie Mannor</p>
<p><u>Memory Intensive Architecture:</u> New memory based machine</p>	<p><i>New Technology based Architecture</i></p>		<p>Dr. Shahar Kvatinsky Prof. Avinoam Kolodny Prof. Eby Friedman (Technion/Rochester) Prof Yuval Cassuto</p>

Ongoing Interaction and Collaboration

Goal: Break-through performance and energy-efficient analytics platform

<u>Time</u>	<u>Plan</u>	<u>Activities</u>
Q1'15	Education, background, select target & workloads.	Bi-weekly * Education * Learning
Q2'15	Broad-stroke microarchitecture, performance, and workloads	* 5/11/15 at Intel * 6/11/15 at Intel
Q3'15	Next-level of detail for microarchitecture, performance, and workloads	Meetings: * DDIO system performance
Q4'15	High-level simulations and models of workloads on microarchitecture	Meetings: * 12/16/16 Update on Weiser Funnel Research
Q1'16	Detailed simulations and models of workloads on microarchitecture	Meetings: * 2/17/16 Update on Ran Ginosaur Accelerators for Big Data Machine Learning * 3/9/16 Update on Shahar Kvatinsky Memory-Intensive Architecture
Q2'16	Detailed simulations and models of workloads on microarchitecture	Meetings: * 4/12/16 Update on Yoav Etsion Prefetcher using NN * 5/2/16 F2F at the Technion
Q3'16 - Q2'17	Parallel work on accelerators for target workloads and microarchitecture.	

Maybe not “detailed simulations and models”
Use the words from the researchers’ stated “major accomplishments”