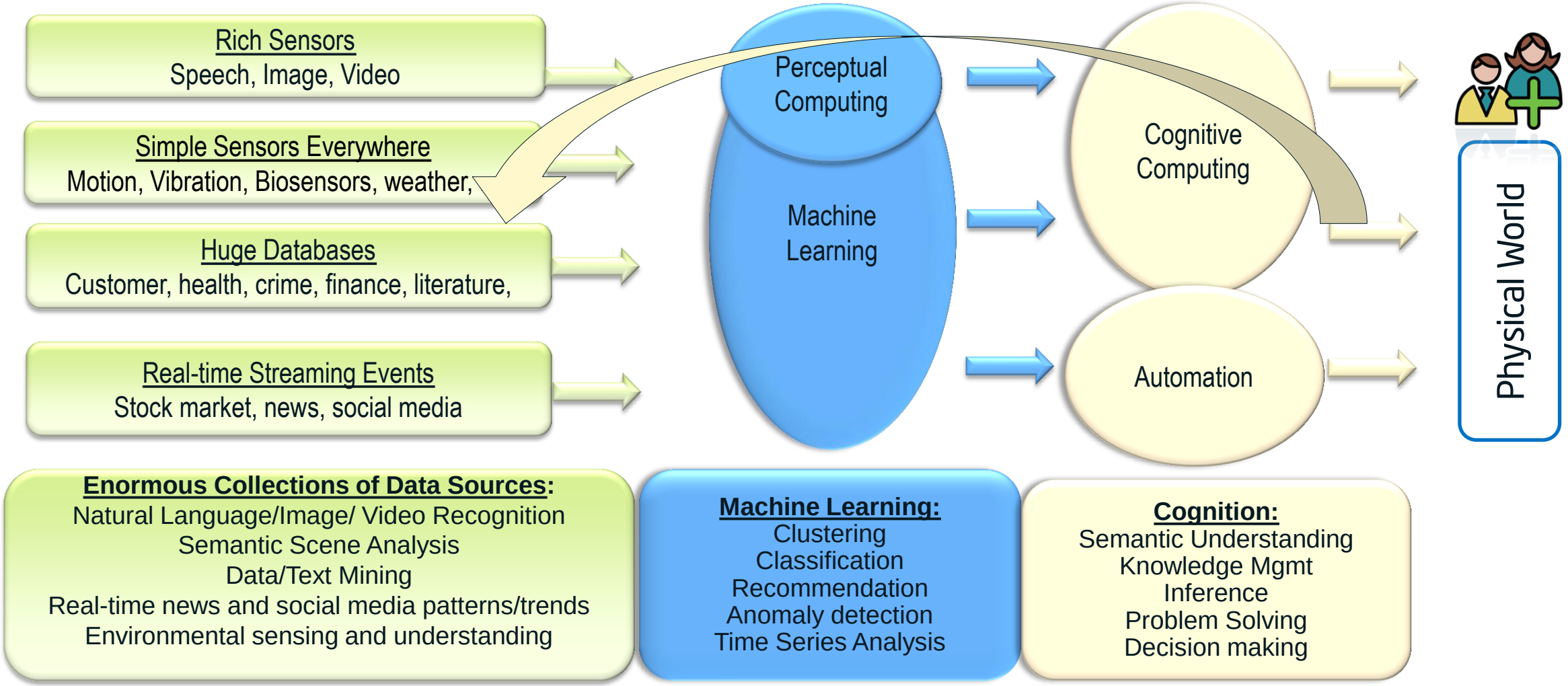
Services

Intel Accelerators View

Debbie Marr & Ravi Iyer

A new Era: Machine Learning, Understanding, Cognition



Explicit Programming → Programming by example, training, tuning

Machine Learning, Visual Understanding, Cognition

→ New, Novel, Effective Platforms

Machines to improve people's lives

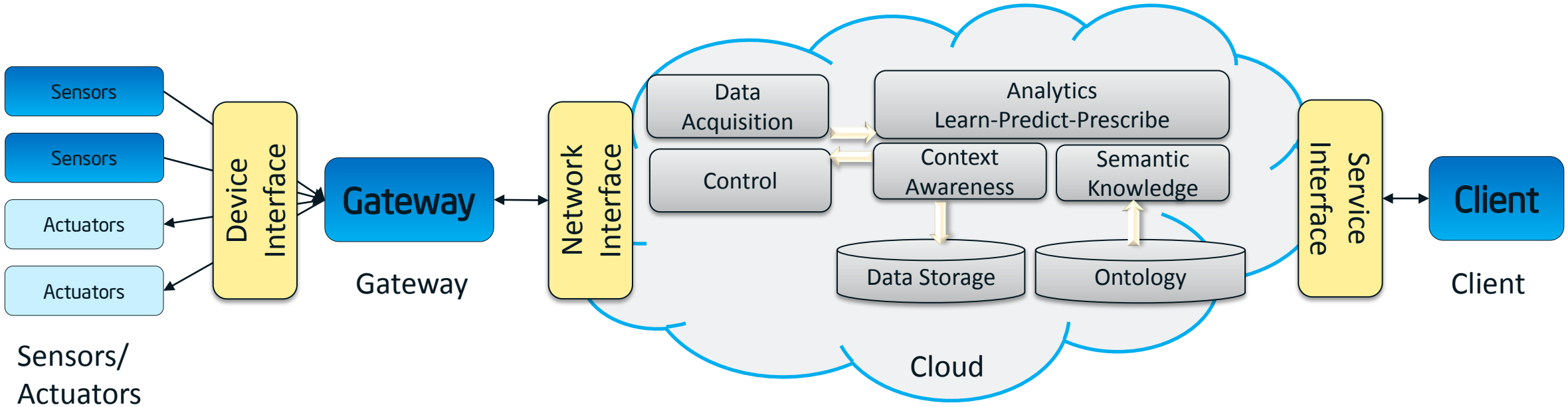
- Internet of Things
  - Understand information
  - Act on the physical world
  - Provide assistance to hu
- Wearables
  - Understand the individual person
  - Active assistance for the individual person
- Work
  - Understanding, cognition, to hel le be most effective to get "work" do
    - Sales & marketing, finance, e udent…
- Cloud Services
  - Combine, synthesize, underst cture
  - Provide services to all

Targeted Accelerators

Accelerators for classes of algorithms

Acknowledgement:
Ramesh Illikkal

Machine Learning, Visual Understanding, Cognition from End to End

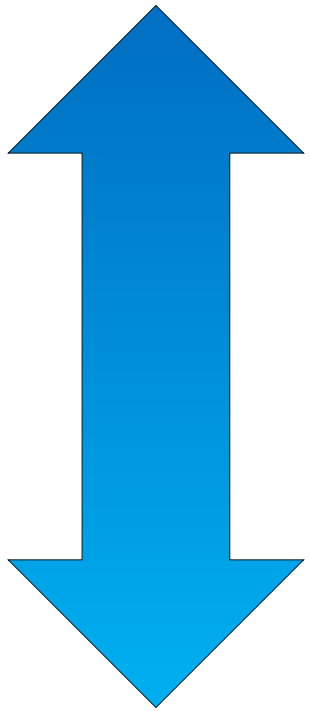


Device Domain | Network Domain | Services Domain | App Domain

Acknowledgement:
Ramesh Illikkal

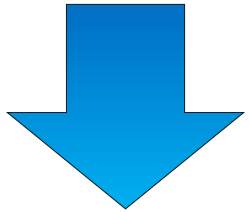
Compute Gap: Explicit programming → Programming by Example

Programming by Example



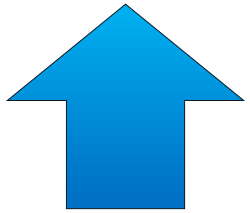
Widening the gap:

- More data
- Improving Accuracy
- Automatic tuning
- Automatic feature selection
- Larger, more complex models



Closing the gap:

- More efficient algorithms
- Accelerators



Explicit Programming

It is the time for Accelerators!

Goal: If it computes, it computes best on Intel platforms

We are:

- Evaluating existing and new algorithm & accelerator ideas
- Optimizing workloads on our platforms
- Prototyping new Wearable & IoT SoC Platforms with sensors, recognition, understanding
- Analyzing Bandwidth, Compute, Scalability, Distributability of these workloads
- Experimenting with new accelerator ideas

Questions for you:

- How are workloads changing? New algorithms, new cognition ideas?
- What are new technologies, new ideas to accelerate these compute problems?

Fine-to coarse-grain acceleration

Continuum of “acceleration”

- Fine-grain
- Medium-grain
- Coarse-grain

Workloads have varying levels of off-loadable work

Overhead

- Performance: Sending work, receiving results, communications/data sharing
- Power: Overhead and power-up/down, sending/receiving/sharing
- Difficult problem: Too often the overhead swamps the benefit!

Compute vs. Bandwidth: When offloading works

Xeon offload to Accelerator on PCIe 2.0 Accelerator Assumed to be 10x Xeon Computation vs. Bytes Moved								
Data Bytes Moved	Xeon Cycles							
	100,000	200,000	400,000	800,000	1,600,000	3,200,000	6,400,000	12,800,000
4,096	4.3	6.0	7.5	8.6	9.2	9.6	9.8	9.9
8,192	4.0	5.7	7.3	8.4	9.1	9.6	9.8	9.9
16,384	3.5	5.2	6.9	8.1	9.0	9.5	9.7	9.9
32,768	2.9	4.5	6.2	7.6	8.7	9.3	9.6	9.8
65,536	2.1	3.5	5.1	6.8	8.1	8.9	9.4	9.7
131,072	1.3	2.4	3.8	5.6	7.1	8.3	9.1	9.5
262,144	0.8	1.5	2.6	4.1	5.8	7.3	8.5	9.2
524,288	0.4	0.8	1.5	2.7	4.2	5.9	7.4	8.5
1,048,576	0.2	0.4	0.8	1.6	2.7	4.3	6.0	7.5
2,097,152	0.1	0.2	0.4	0.9	1.6	2.7	4.3	6.0
4,194,304	0.1	0.1	0.2	0.5	0.9	1.6	2.7	4.3
8,388,608	0.0	0.1	0.1	0.2	0.5	0.9	1.6	2.8

Benefit of offloading to an accelerator assuming ~10,000 cycles of overhead

Analytic model, can trade-off overhead & data bandwidth vs. accelerator speedup

Make it easy for software developers!

Unreasonable for software to use arbitrary accelerators

Need simple model for sw developers

Accelerator offloading done under-the-hood

Ninja programmers developing a single application which runs alone on a fixed platform is the exception, not the rule!

Your work is very relevant to the new era in computing

Advanced Machine Learning

- Fundamentals & improvements to the learning algorithms. Scalability & Distributability of algorithms. Optimizing the "teacher's" time. NLP Parsing.

Imaging & Visual Understanding

- Ofri already talked about this yesterday. Algorithms, usages, and accelerators! Really important for IoT, wearables, pushing to the edges. Power. Accelerators.

New Accelerator and New Microarchitectures

- Power, Energy, Performance, Accelerators, New novel microarchitectures

Cognition

- New usages, new algorithms, new computing paradigms.

Summary

A new era of computing is coming

It brings a new compute paradigm: Programming by example

New architectures and microarchitectures are needed - from end-to-end