# Real-time Machine Learning

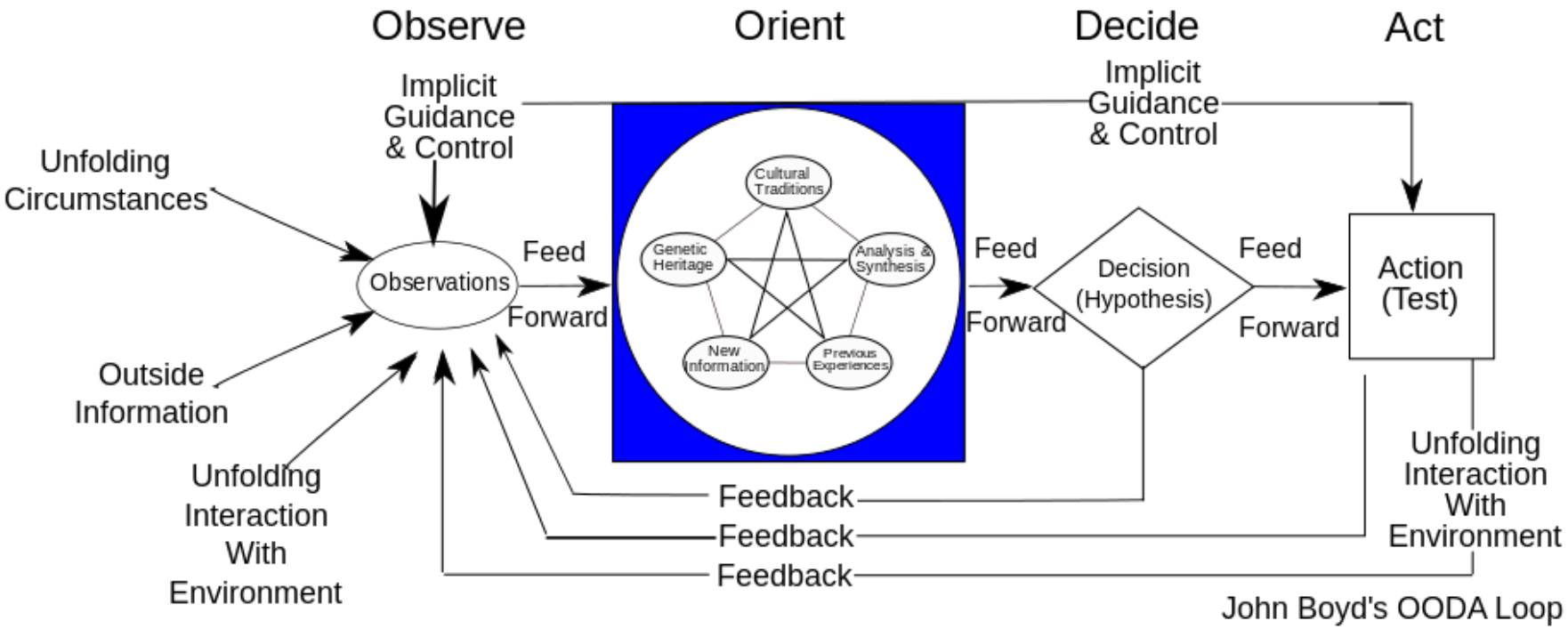
## Big Ideas for a Mega City (actually Capability Concepts)

Philip Leong | Computer Engineering Laboratory
School of Electrical and Information Engineering,
The University of Sydney

THE UNIVERSITY OF SYDNEY

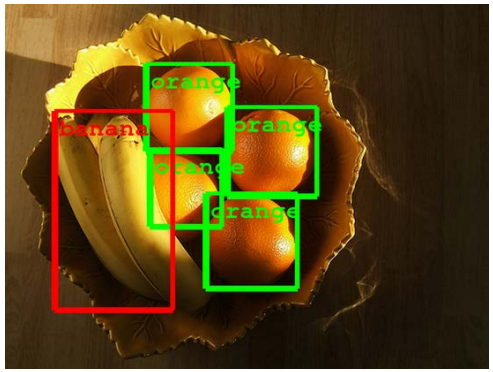
# Decision making: OODA Loop

Explicit process used to thrive in an ambiguous environment




**Applies to war, business or life.**

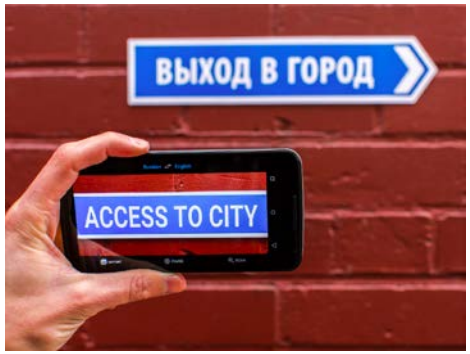
# Contemporary Machine Learning (ML)




Computer vision

Speech recognition

Natural language processing




Machine translation

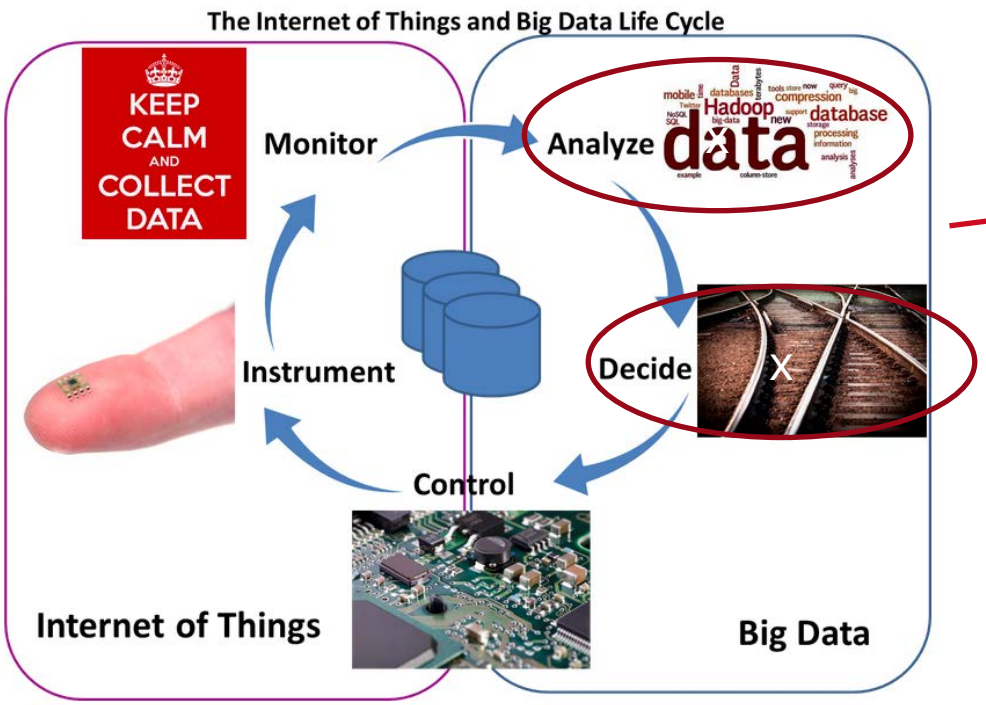
Recommender systems

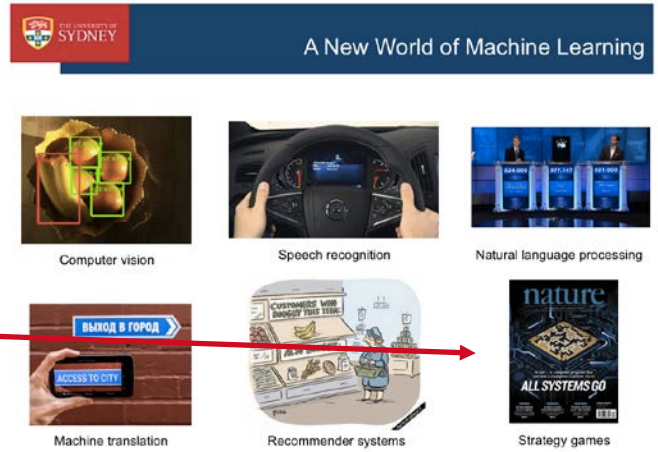
Strategy games

Source: Google

# Contemporary IoT

- We can already use OODA e.g. in high-latency smart-grid applications
  - e.g. we cannot prevent power outages due to cascading faults



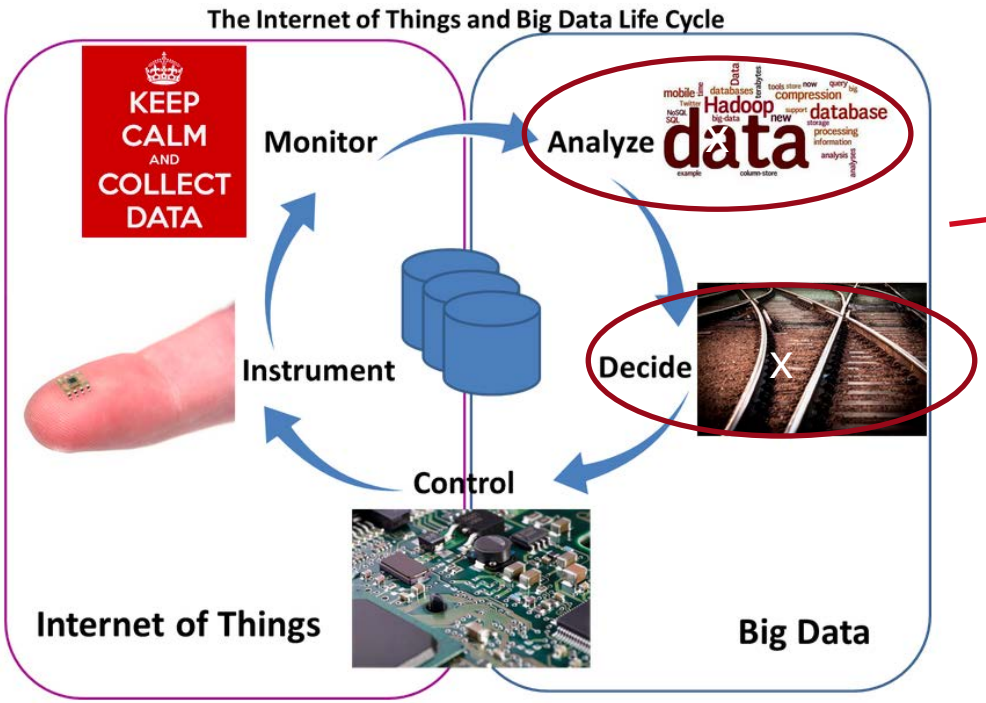

Source: Harish Vadada

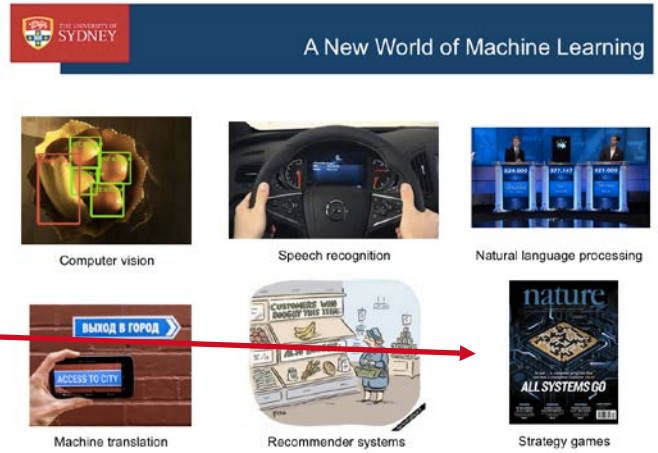
# Missing Link

## We cannot go around OODA loop in a short time (say 100 ns)

› Only part requiring disruptive technology is low-latency machine learning

- Must be Energy efficient, Parallel, Integrated, Customisable




- A single, embedded device that can do all this - **EPIC chip**

# Wouldn't it be expensive to develop?

- Not at all, the hardware and systems are available (FPGAs)
  - ARM processor
  - 12-bit ADCs @ 4 Gsps
  - 14-bit DACs @ 6.4 Gsps
  - FPGA fabric
  - Memory
  - High speed transceivers
  - Interfaces to anything
  - Low power
  - Local industry

- Can be reprogrammed to change functionality and update algorithms
  - Could do real-time voice recognition and face recognition on the same device

Source: Xilinx

# Applications

Applications range from simple and clearly feasible, to futuristic

- Prognostics
  - Monitor machines or systems and take action before damage is done (e.g. prevent power outages due to cascading faults)
- Smart City
  - Sensing and surveillance (e.g. intelligent traffic lights, monitor potential terrorists)
- Signal processing
  - Understanding radar, sonar, HSI data (e.g. auto maneuver to avoid torpedo)
- Robotics
  - Improve situational awareness (e.g. small autonomous drones conducting missions)
- Cybersecurity and electronic warfare
  - Anomaly detection and information understanding at network speeds (e.g. respond to jamming in a timely fashion)

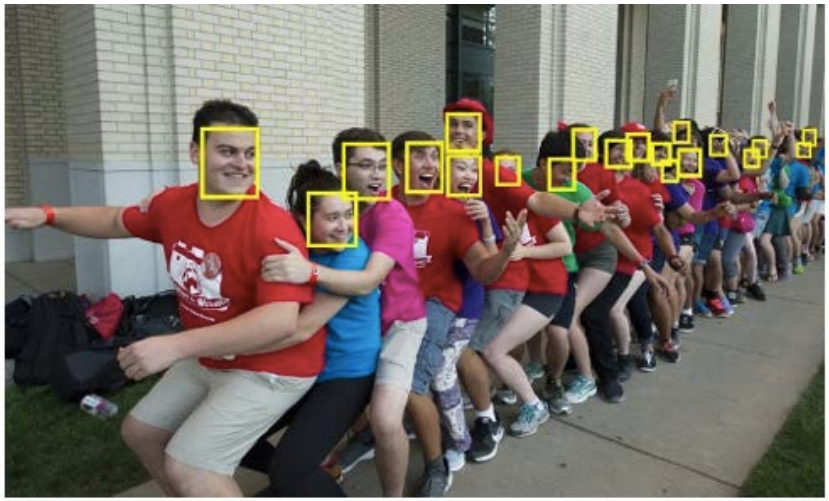
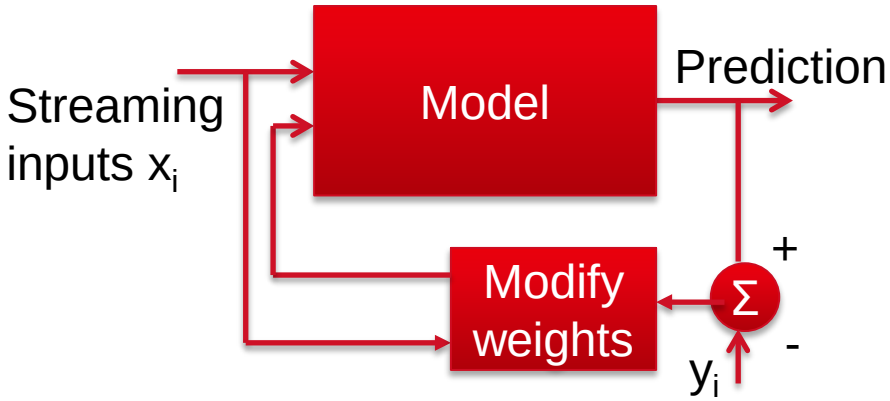
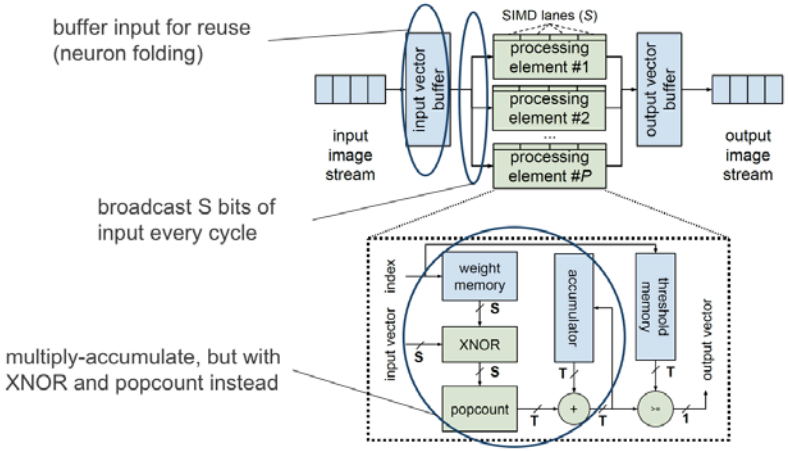


Photo: CMU

# Our Recent Research

We have developed following low-latency FPGA-based components



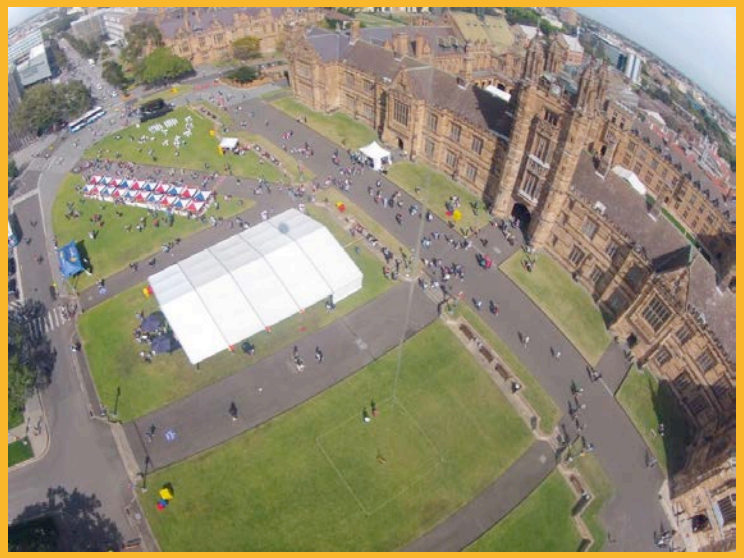




**Kernel methods with state-of-the-art latency (< 100ns) and throughput (70 Gbps)**

**Binarized convolutional neural networks with state-of-the-art performance (25W, 12.3 Mcps, 0.31 us latency, MNIST 95.8% accuracy)**

# Summary

- New technologies
  - ML advances allow Orient and Decide parts to be better than humans
  - FPGAs allow an OODA loop to execute with unprecedented speed
- Idea
  - Create capability which combine the two
  - Develop the necessary ML theory and implementations (EPIC chip)
- Certainly emerging and disruptive technology with widespread applications
  - Everybody investing in large-scale ML but little in low-latency



Thank you!

**Philip Leong (philip.leong@sydney.edu.au)**
**http://www.ee.usyd.edu.au/~phwl**