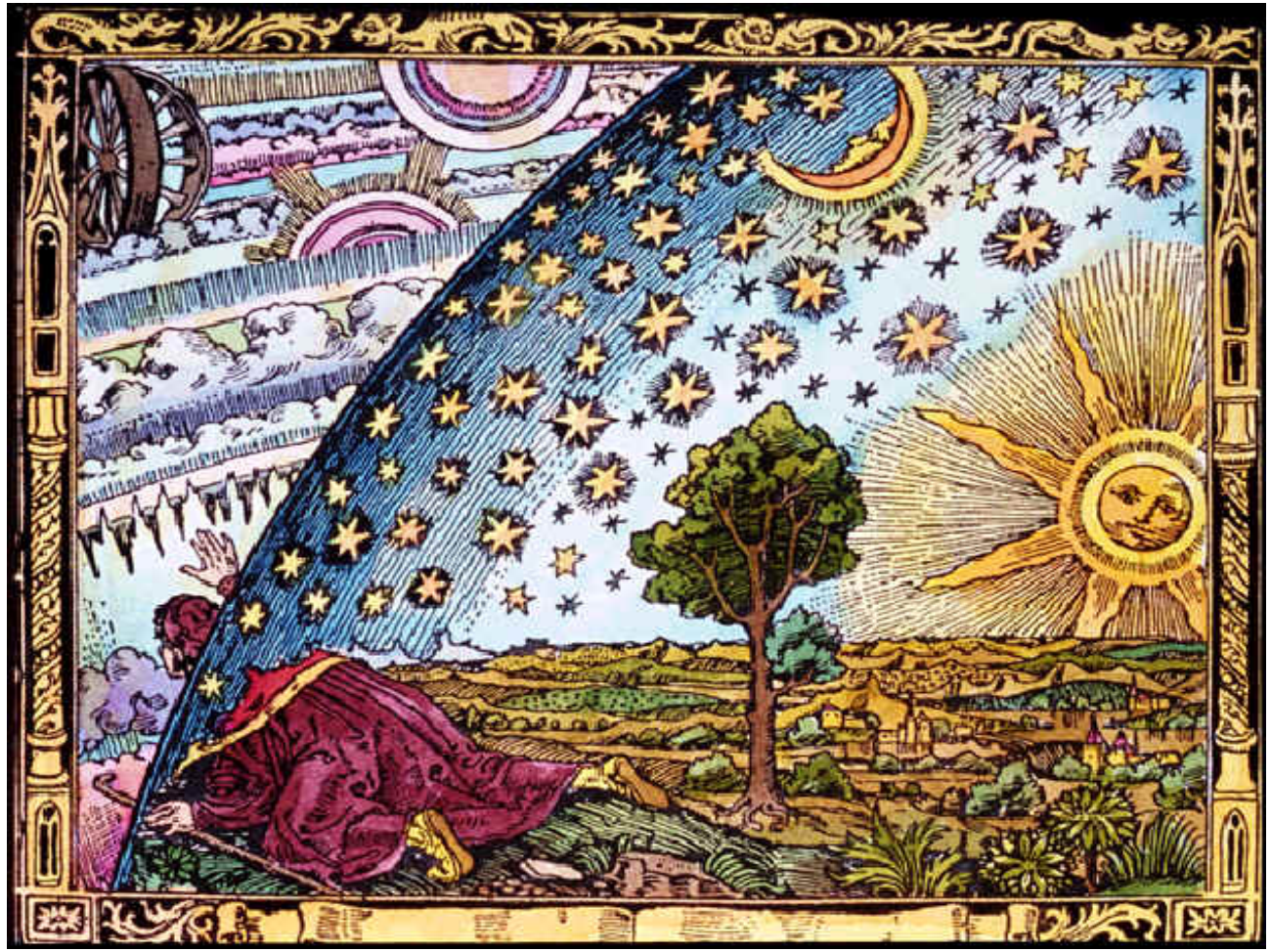
Computing at the Edge:
Sensors, Learning, and Adaptation



Dan Reed
University of Utah

dan.reed@utah.edu
www.hpcdan.org

THE UNIVERSITY OF UTAH®

Quantitative creates qualitative …
… and it changes everything

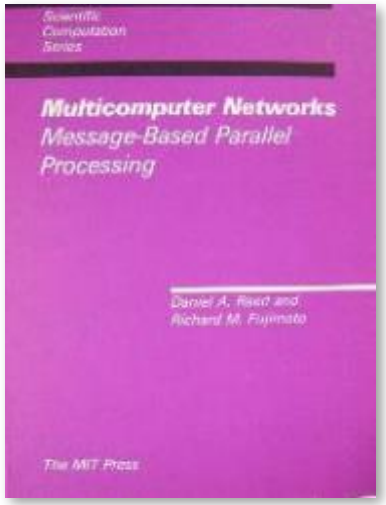
Viability determined by

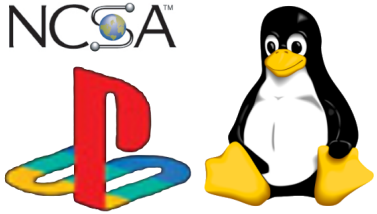
- *speed, capacity, cost, market scale*

… and their ratios

NCSA Cray X-MP (1985)

$8,000,000 and ***56 Kb/s NSFnet***
800 megaflops (peak)
(~$18M in 2018 dollars)





TeraGrid™

Intel 80386 (1985)
16 MHz @ 2.3 watts
1.5 micron CMOS

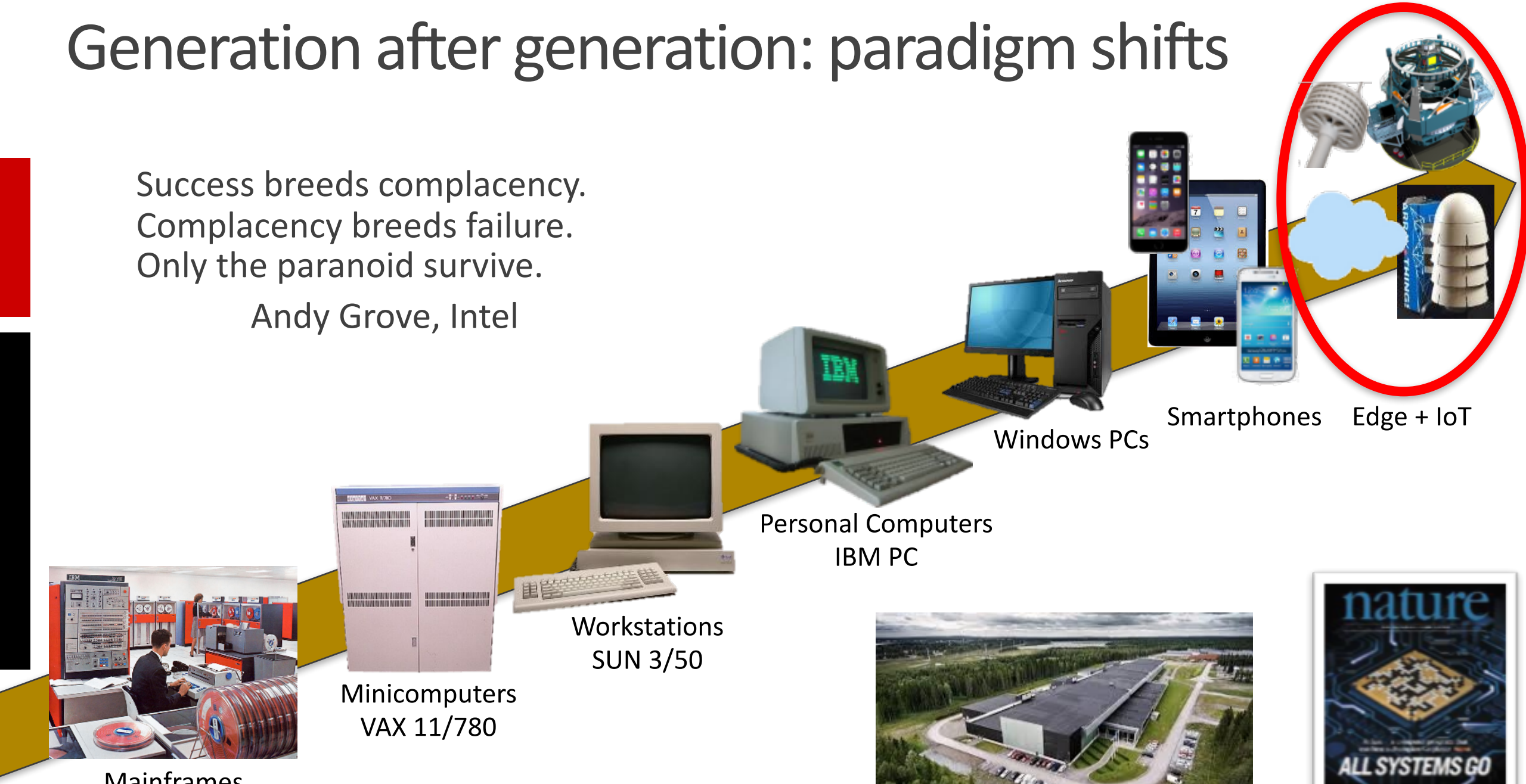
iPhone 8 (2017)
~$700 and LTE wireless
~3000 megaflops

THE UNIVERSITY OF UTAH®

Generation after generation: paradigm shifts

Success breeds complacency.
Complacency breeds failure.
Only the paranoid survive.

Andy Grove, Intel



Mainframes
IBM S/360

Minicomputers
VAX 11/780

Workstations
SUN 3/50

Personal Computers
IBM PC

Windows PCs

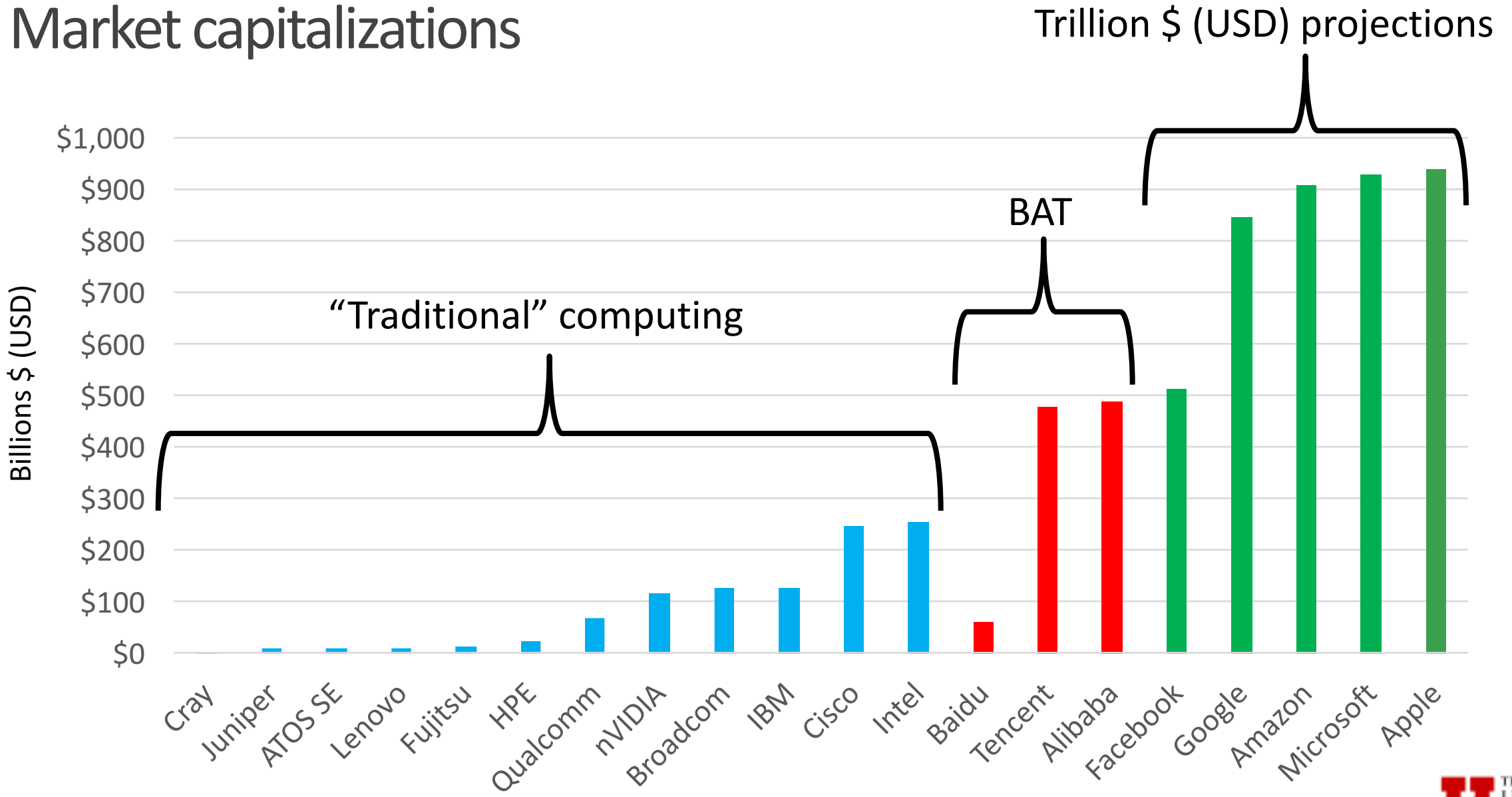
Smartphones

Edge + IoT

Cloud Data Centers

Our world has changed

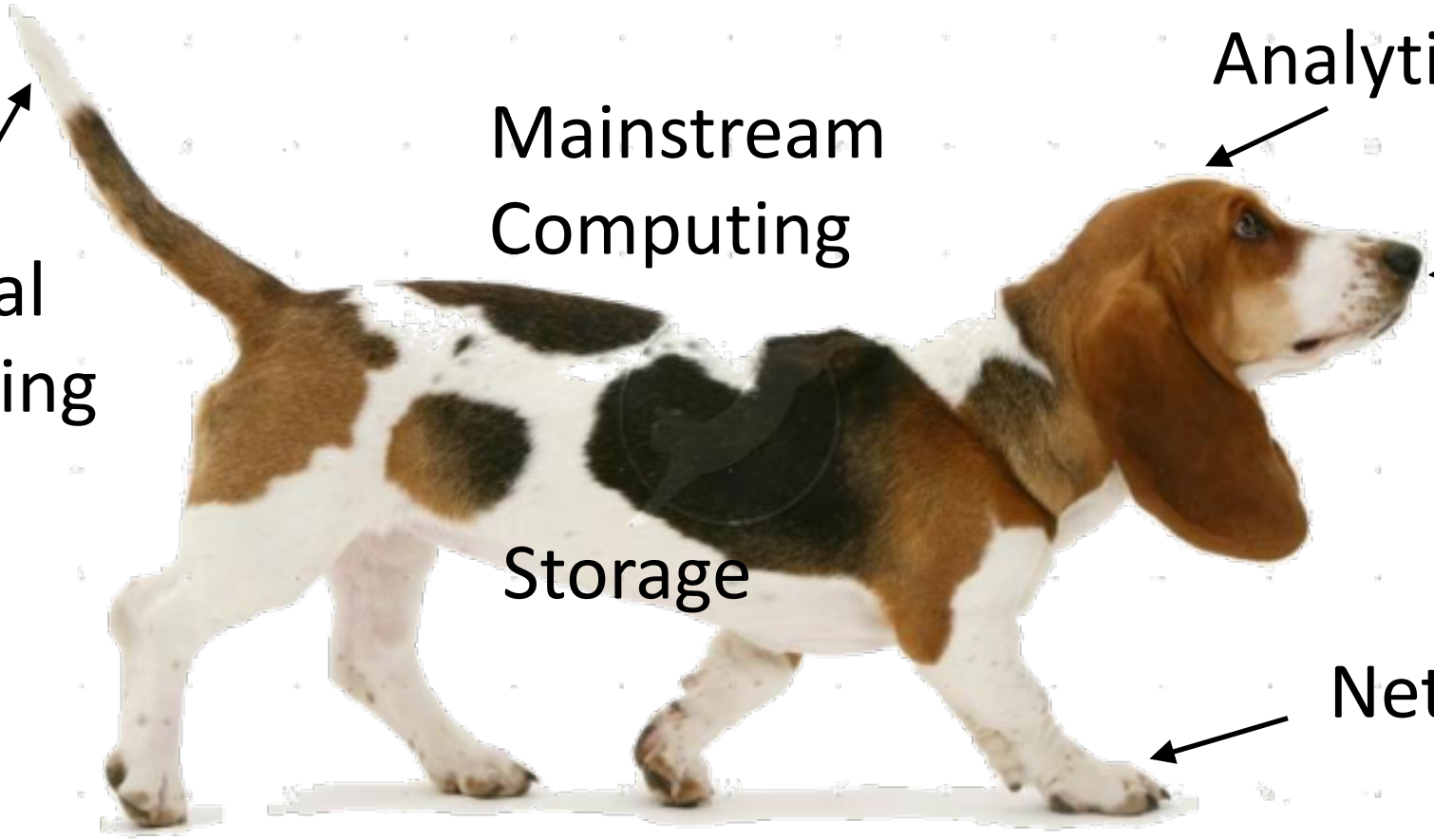
Market capitalizations



April 2019

THE UNIVERSITY OF UTAH

Follow the money and the users …



From petascale to the global race to exascale

ORNL Summit (200 PF)

From K to Post-K Computer

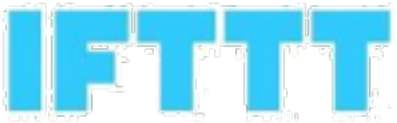
THE UNIVERSITY OF UTAH

Meanwhile, everything gets smart …

… and the trolley paradox gets real

Oura Ring

Oxford Nanopore
Mobile DNA sequencer



THE UNIVERSITY OF UTAH

The computing continuum: holistic thinking needed

Edge — **Fog** — **HPC/Cloud/Instrument**

Size	Nano	Micro	Milli	Server	Fog	Campus	Facility
Example	Adafruit Trinket	Particle.io Boron	Array of Things	Linux Box	Co-located Blades	1000-node cluster	Datacenter & Exascale
Memory	0.5 KB	256 KB	8 GB	32 GB	256 GB	32 TB	16 PB
Network	BLE	WiFi/LTE	WiFi/LTE	1 GigE	10GigE	40GigE	N*100GigE
Cost	$5	$30	$600	$3K	$50K	$2M	$1000M

Count = 10^9
Size = 10^1

Count = 10^1
Size = 10^9

Count x Complexity = ~Constant

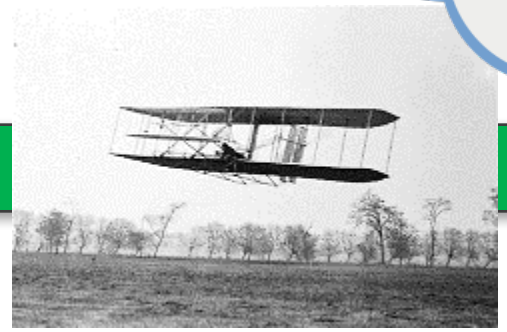
Stateful vs. Stateless

Source: Beckman, Beck, Dongarra, Ferrier, Reed, and Taylor

Three computing revolutions …

Performance

Deep Learning



Edge & AI

Google Cloud

amazon web services™

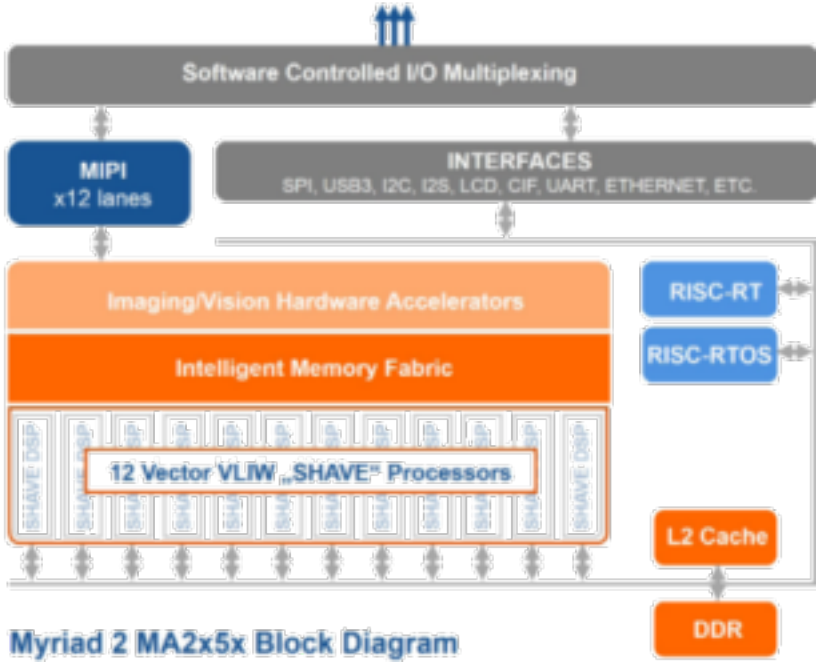
Microsoft Azure

THE UNIVERSITY OF UTAH®

Deep neural networks at the edge

Most data *never* leaves the device

$249 (USD)



Myriad 2 MA2x5x Block Diagram

~$80 (USD)

Amazon Deep Lens

- Integrated camera
- Intel Atom and Ubuntu 16.04
- Intel Gen9 graphics engine
- TensorFlow and Caffe support
- AWS integration (obviously)

Movidius (Intel) Neural Compute Stick

- Custom Vision Processing Unit (VPU)
- TensorFlow and Caffe support

THE UNIVERSITY OF UTAH®

Edge ferment

Coral USB

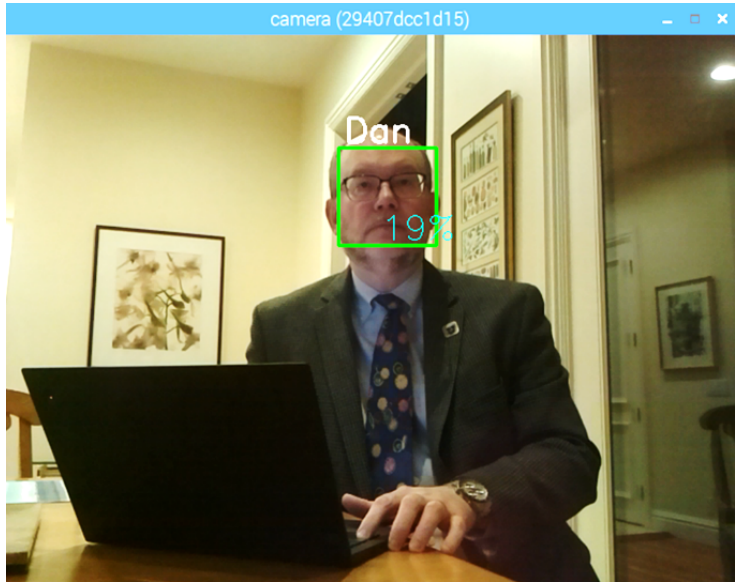
Google Coral ($149)

- Quad-core Cortex-A53 + Cortex-M4F
- *Google Edge TMP ML accelerator*
- 8 GB eMMC
- 1 GB LPDDR4
- GigE, HDMI, USB-C, USB-3
- MicroSD slot

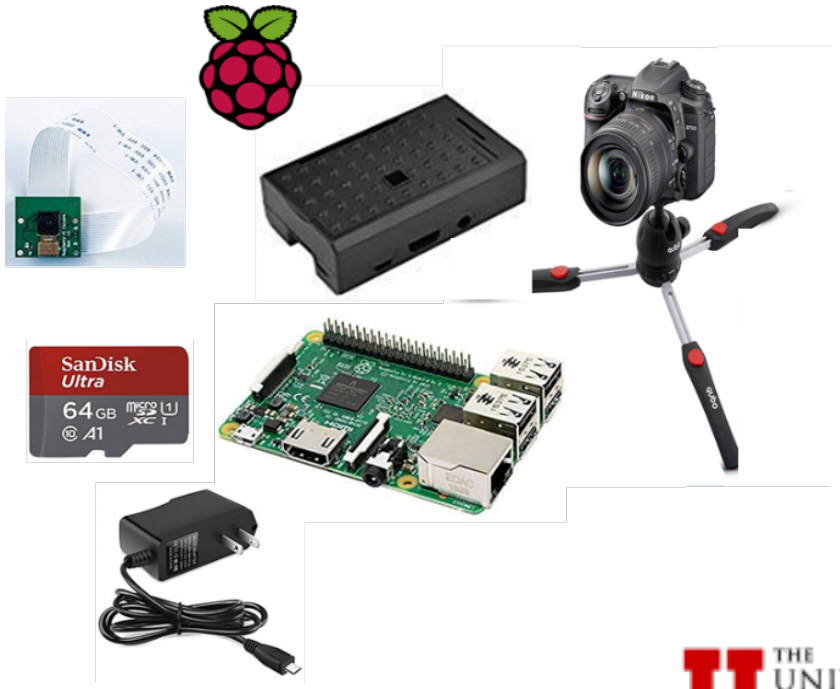
Jetson Nano ($99)

- Quad-core A57
- *128 core nVidia Maxwell*
- 4 GB LPDDR4
- GigE, HDMI, USB 3
- MicroSD slot

Face recognition on the cheap



Hardware Item	Cost
Raspberry Pi 2/3 Case (optional)	$9.99
AZ Delivery Raspberry Pi Camera	$7.99
Raspberry Pi Model 3	$34.99
SanDisk 64 GB microSDXC UHS-I card	$11.59
5V 2.5A Micro USB AC Adapter	$10.99
Qubo Phone Tripod Camera Stand (optional)	$20.99
TOTAL	**$96.54**



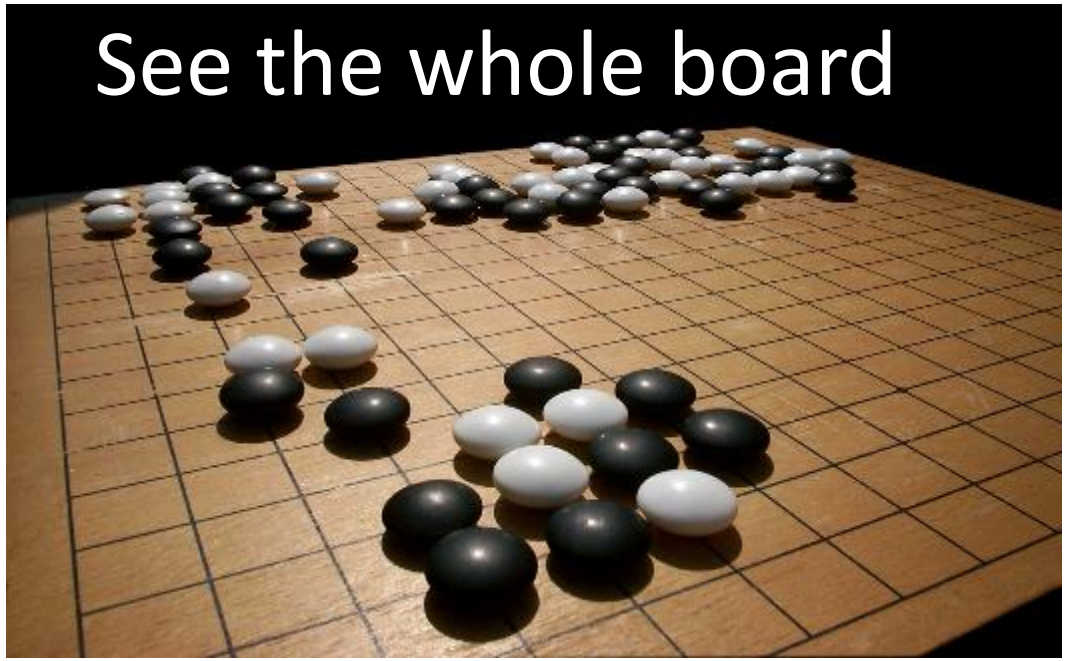
THE UNIVERSITY OF UTAH®

Serious action

Now at the two extremes

The very small (edge/fog computing and sensors)
The very large (clouds, exascale, and big data)

But with real constraints on each



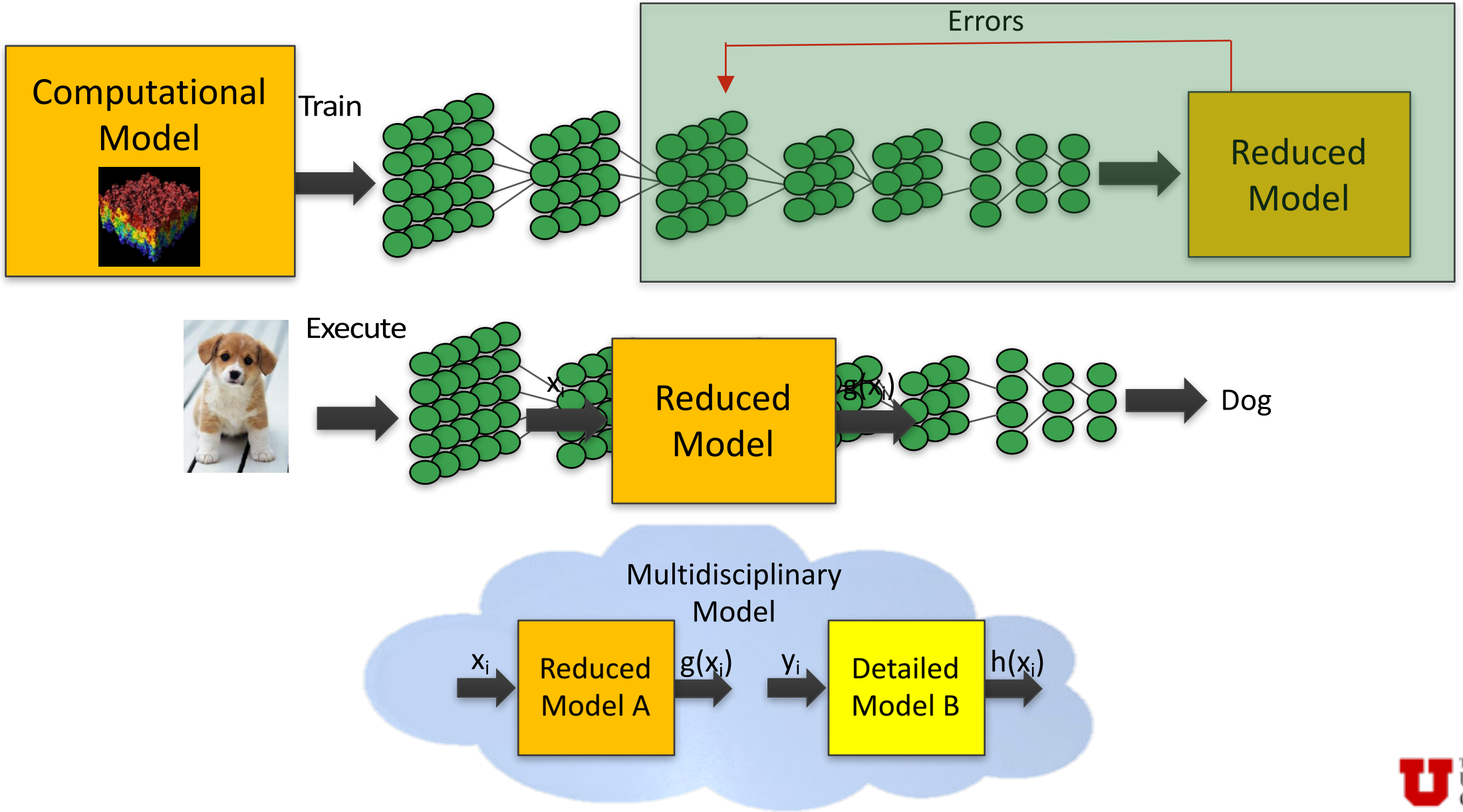
Technical implications

- Fluid end-to-end cyberinfrastructure
- Interdisciplinary data and infrastructure sharing

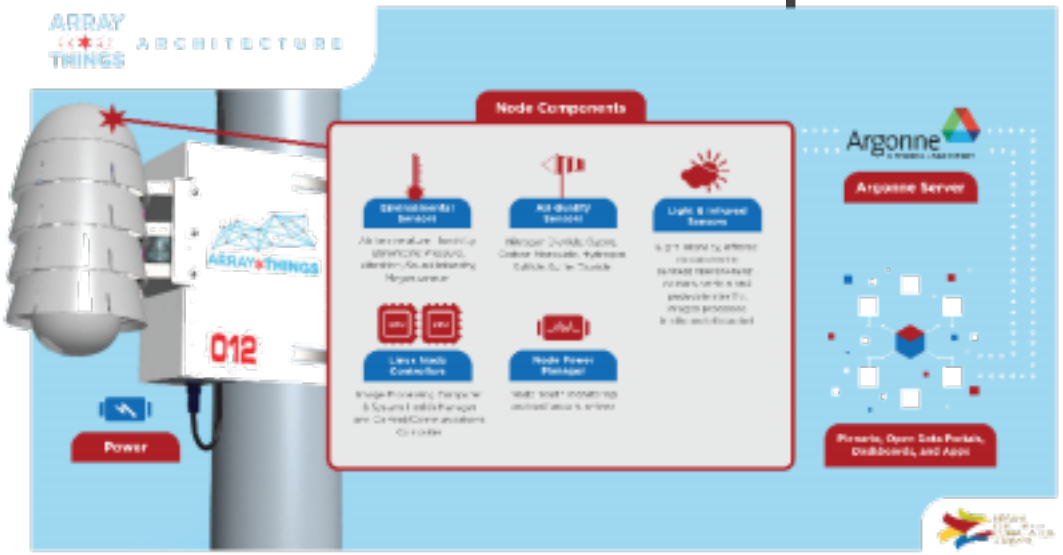
Cultural implications

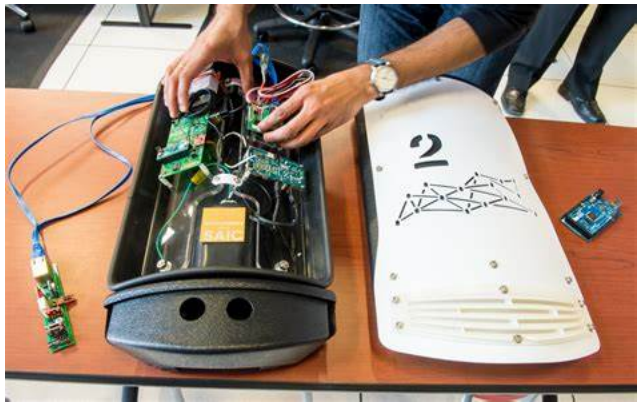
- Change management and strategic planning
- Community collaboration
- “Run forever” applications

Deep nets and complex models at the edge



Smart cities and personalized medicine

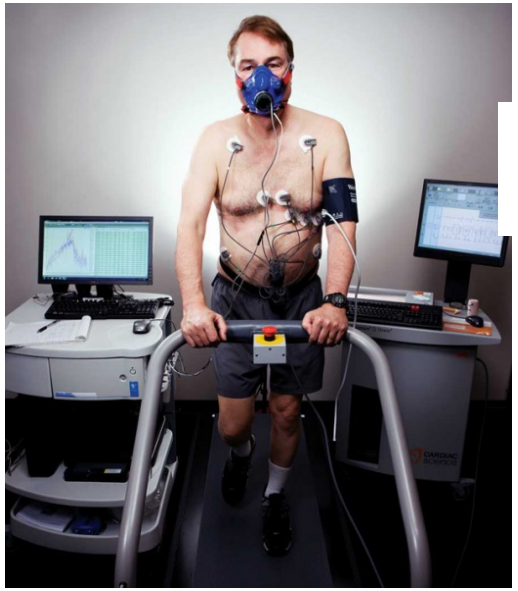




Charlie Catlett

ARRAY OF THINGS

https://arrayofthings.github.io



Oura Ring

Oxford Nanopore
Mobile DNA sequencer

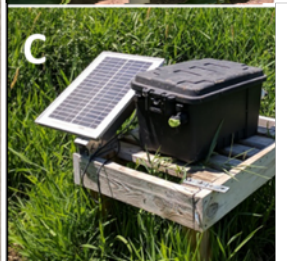
THE UNIVERSITY OF UTAH®

A framing question …

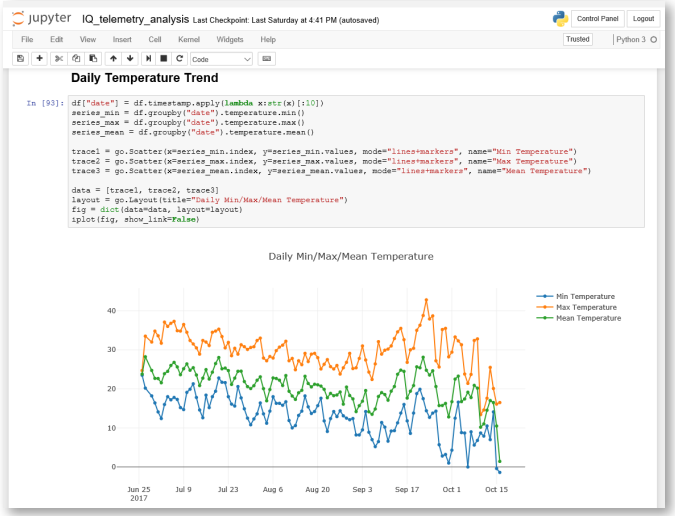
How would you use

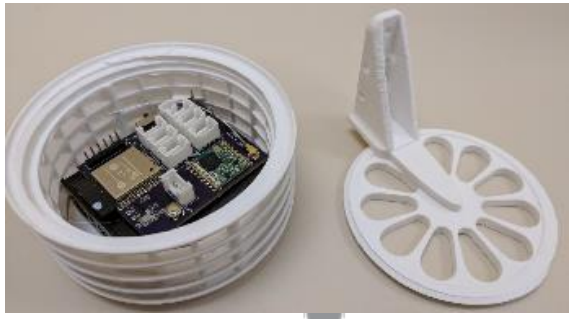
- Hundreds of ~$50 wireless sensors?
- Streaming environmental data

to understand watersheds and irrigation?



Solar powered Raspberry PI gateway





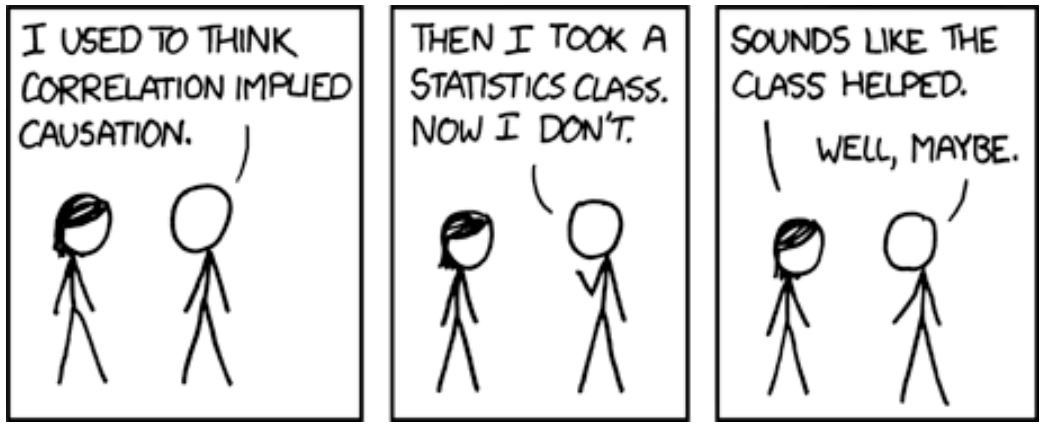
Custom Arduino Sensor

Kang, Kuhl, Bockholt, Rogers, and Reed, "A Cloud-Based Scientific Gateway for Internet of Things Data Analytics," *PEARC18*

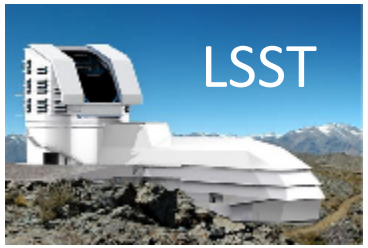
Big data: changing perceptions, shifting challenges

What information consumes is rather obvious: it consumes the attention of its recipients. Hence *a wealth of information creates a poverty of attention*, and a need to allocate that attention efficiently among the overabundance of information sources that might consume it.

Herbert Simon



Science instrument continuum



BIG and FEW

Multidisciplinary Data Fusion & Modeling

SMALL and MANY

Vehicles

Array of Things

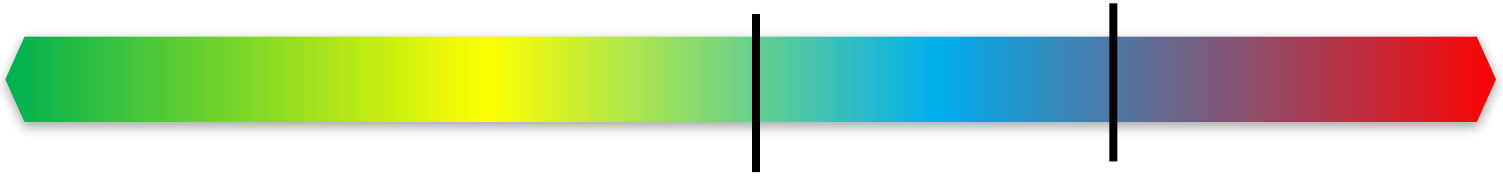
Health

THE UNIVERSITY OF UTAH

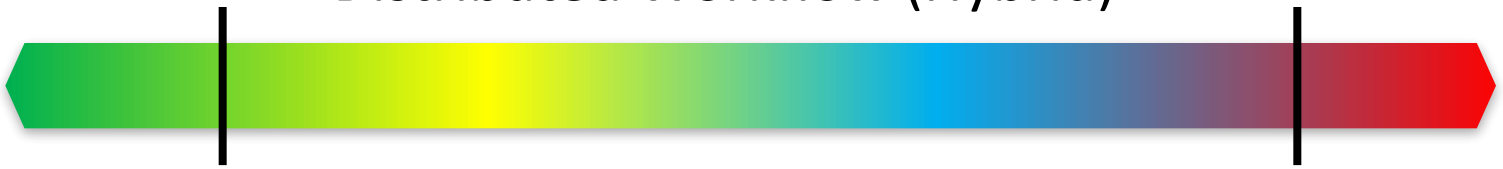
Building fluid capabilities: AND not OR

LHC, LSST, SKA (Big Instrument) and HPC

Smart City/Environmental Monitoring (Edge)



Distributed Workflow (Hybrid)



Stateless

Mapping
- What
- Where
- When

Subject to
- Speed
- Capacity
- Latency
- Resilience
- Security

Stateful

AWS Greengrass

Harnessing the computing continuum

Intentional specification

Science-driven Problems

e.g.: “Predict urban response to rainfall, trigger intelligent reaction…”

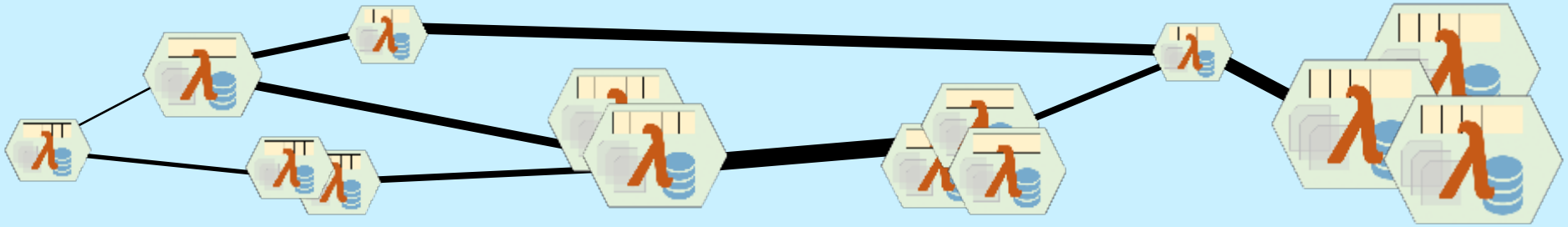
Research Focus

Goal-oriented Annotations

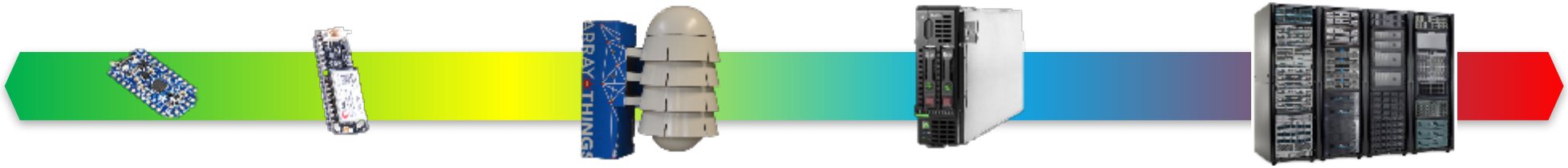
Notional Example:

trigger {flood_actuation, resident_warning}
when {wx_prediction, sewer_model} implies
(traffic_capacity < 70%) or (home_flooding > 5%)

Continuum Abstract Model & Runtime



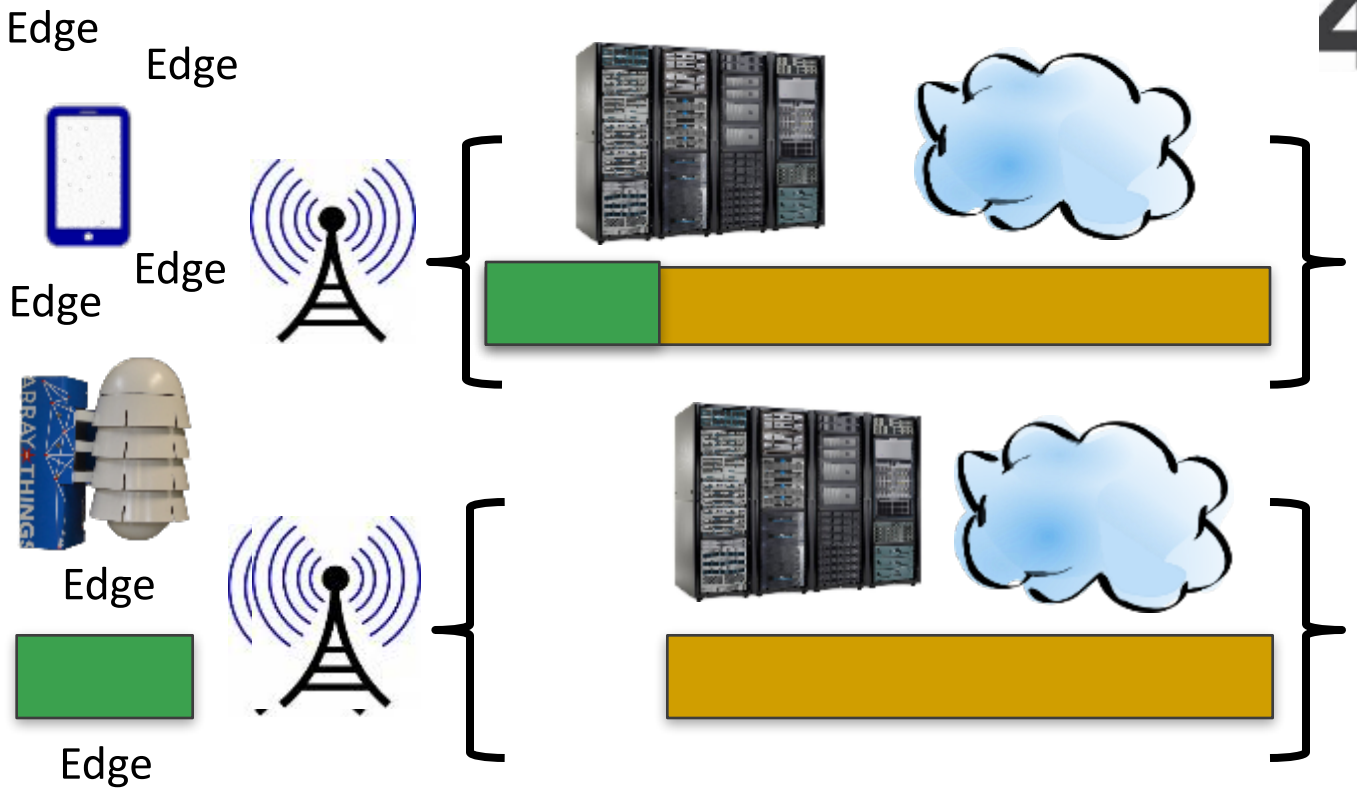
Existing Resources & Services



Source: Beckman, Beck, Dongarra, Ferrier, Reed, and Taylor

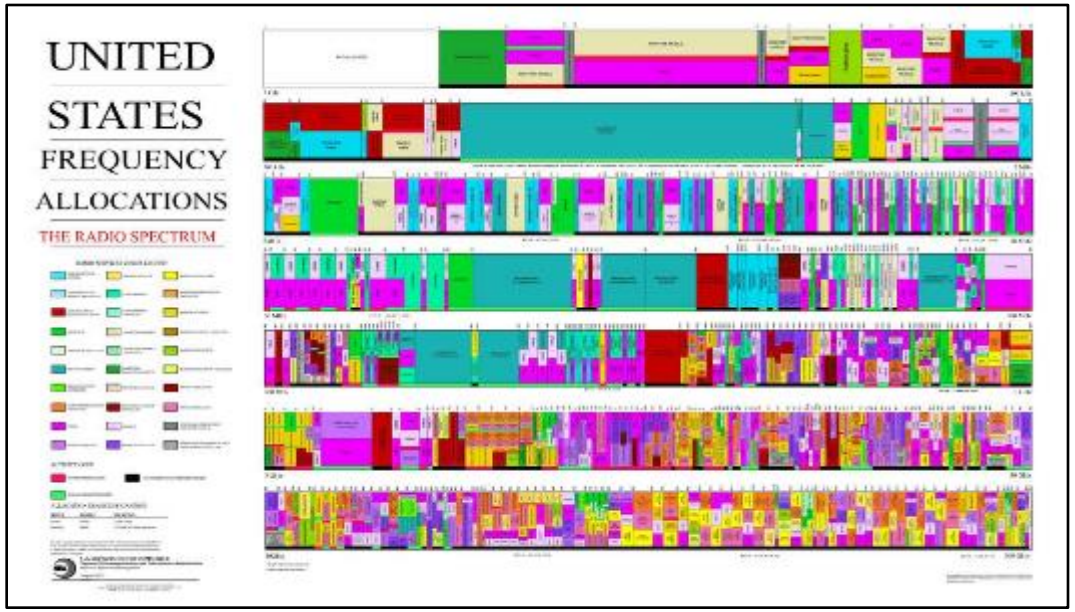
THE UNIVERSITY OF UTAH®

Building fluid capabilities



Latency
Bandwidth
Energy

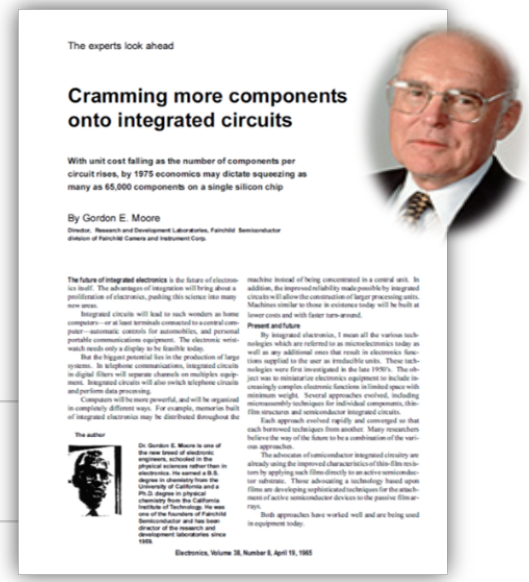
Storage
Knowledge
Context

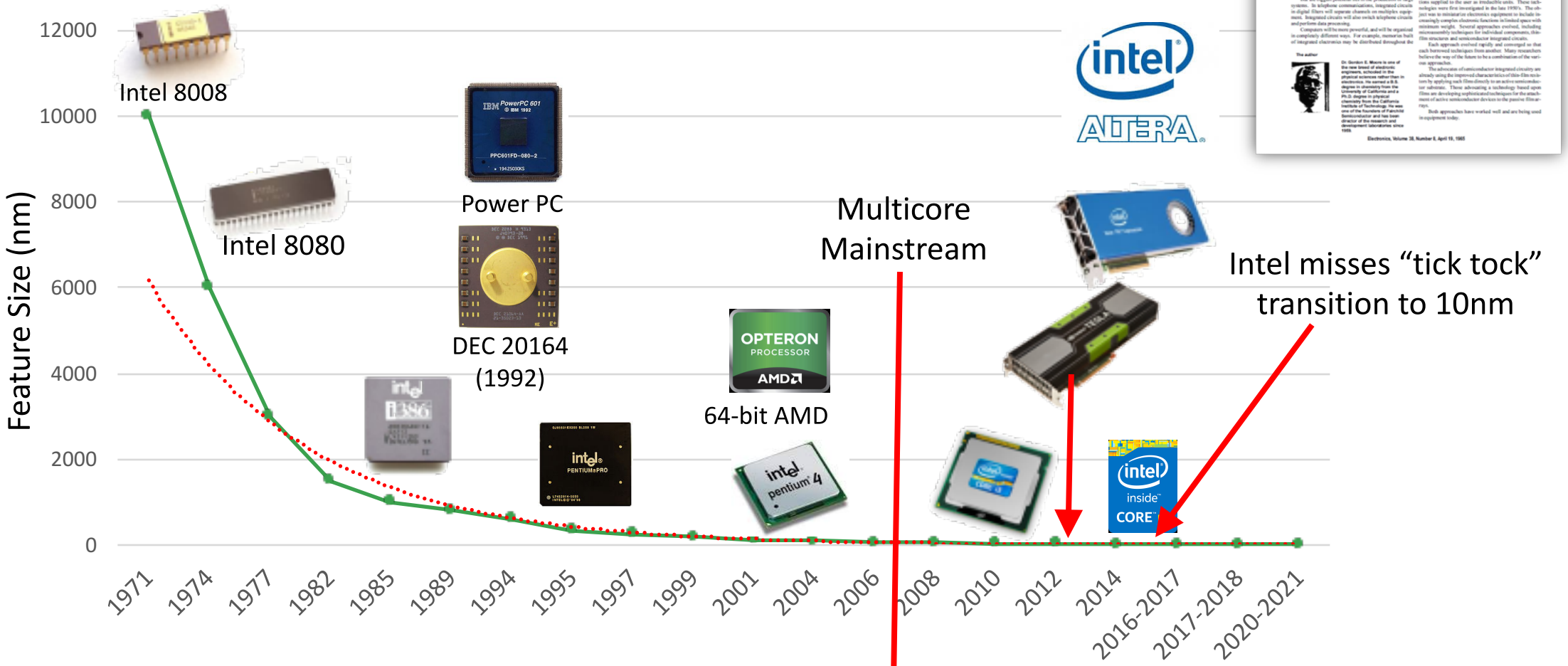


THE UNIVERSITY OF UTAH

Chip feature sizes and Dennard scaling

No exponential is forever (except in the textbooks)





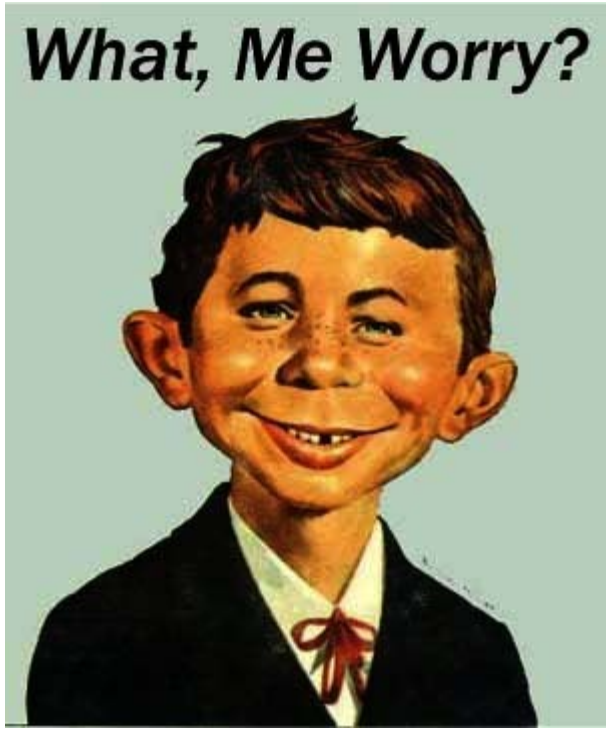
https://science.energy.gov/~/media/ascr/ascac/pdf/meetings/201812/ASCAC_BRN_Microelectronics.pdf
https://science.energy.gov/~/media/bes/pdf/reports/2018/Microelectronics_Brochure.pdf

THE UNIVERSITY OF UTAH

Why us? Why now? (channeling Jim Larus)

Why us?

- Someone needs to think about higher level models



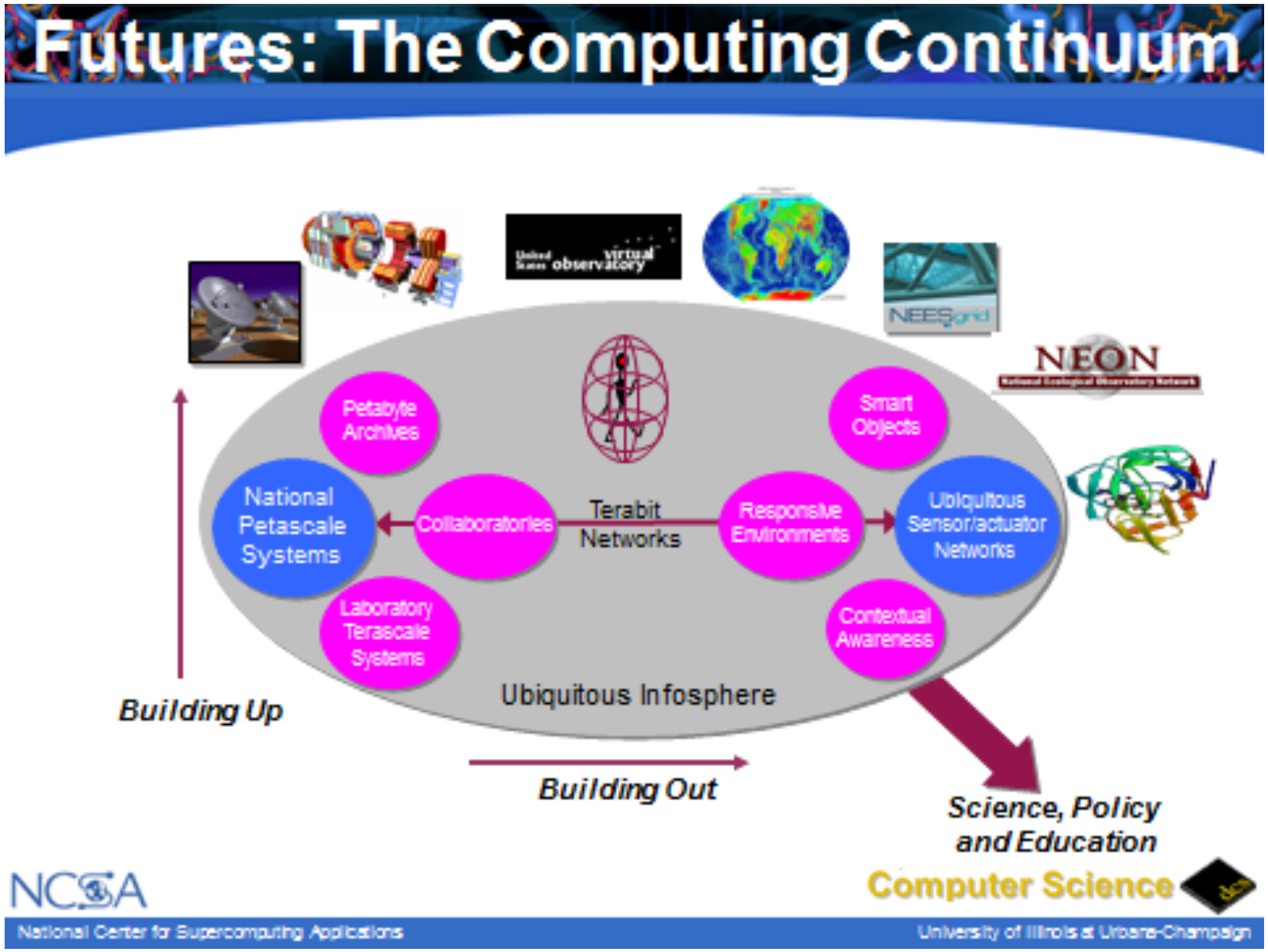
Why now?

- End-to-end, “run forever” services are the future
- Architectural specialization puts a premium on portability
  - Minimizing data movement at many levels
  - Maximizing operations/joule

Why fusion?

- Integration will enable new capabilities
- It’s more than workflows, containers, and libraries

Not yet fully realized … my 2002 prediction



THE UNIVERSITY OF UTAH

The big questions don’t change …

Discussion

… but the approaches and answers do

THE UNIVERSITY OF UTAH®