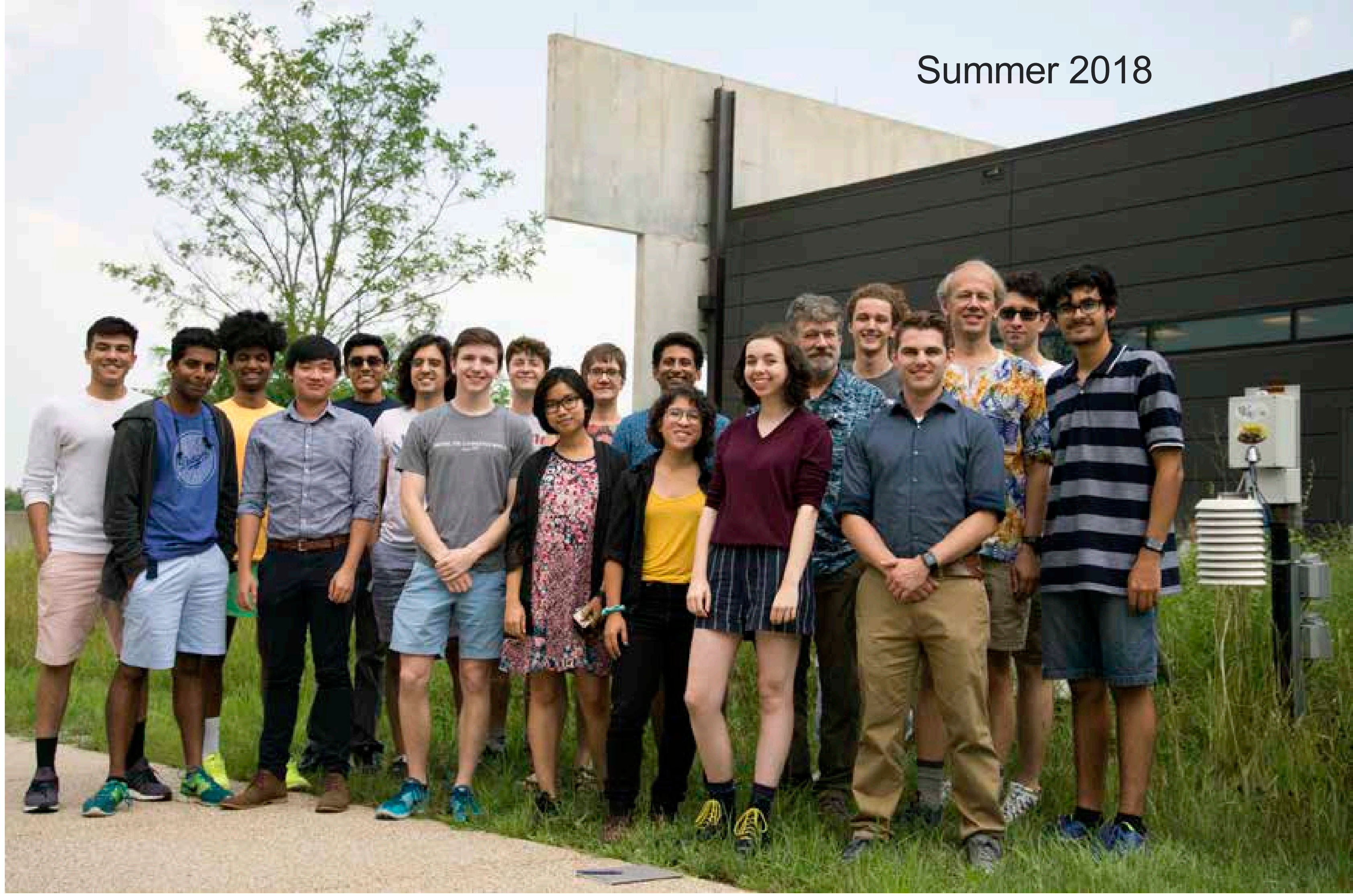
Artificial Intelligence at the Edge



Summer 2018

Pete Beckman, Nicola Ferrier, Charlie Catlett, Rajesh Sankaran

Co-Director Northwestern University / Argonne Institute for Science and Engineering (NAISE)

Argonne National Laboratory, Northwestern University, University of Chicago

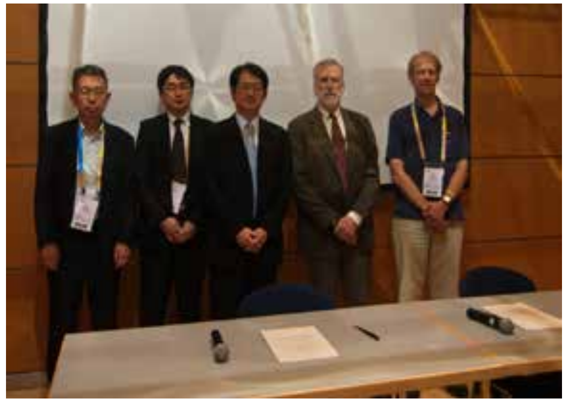
Argonne NATIONAL LABORATORY

Outline

- Stumbling to the Edge
- A Waggle for Rough Edges
- At the Edge of Chicago
- Science on the Edge
- Cutting Edge Hardware
- Edgy Topics for R&D

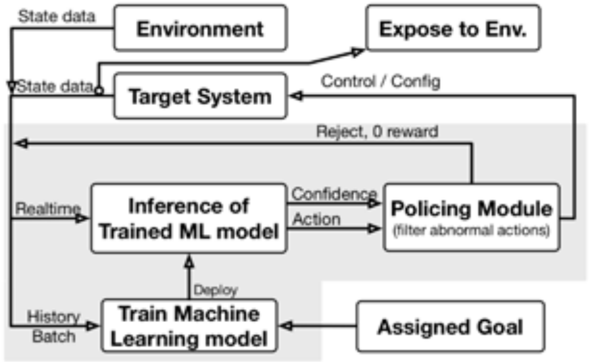
Not Today, but find me if you are interested

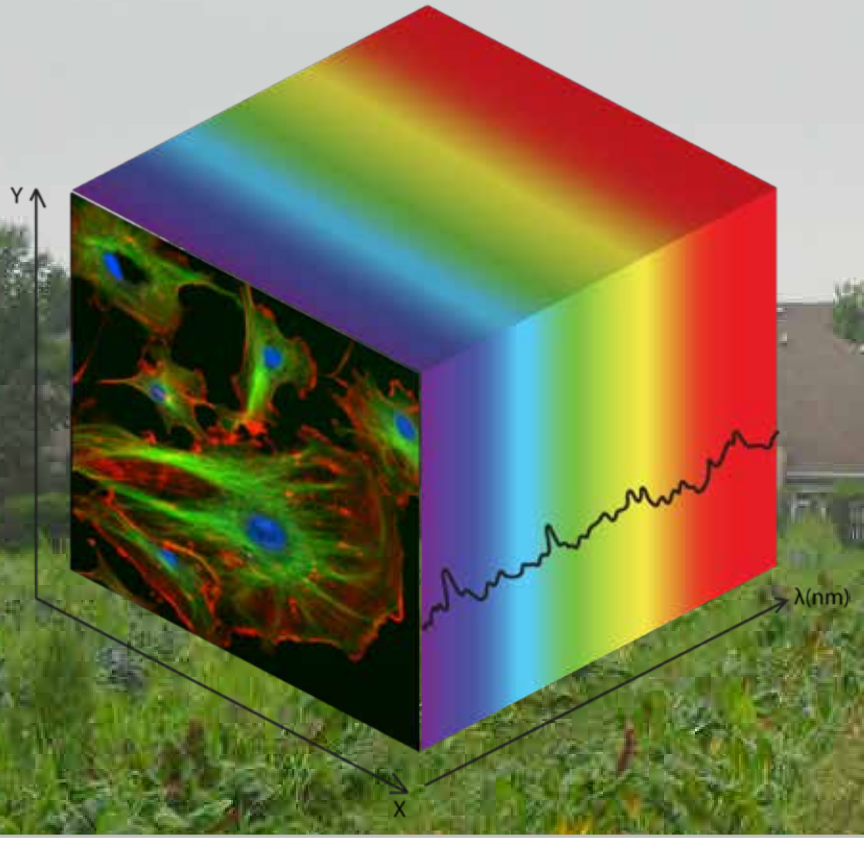
EXASCALE COMPUTING PROJECT



USA-DOE
JP-MEXT

AMASE
Smart
HPC





Example: SPECIM Camera:
PFD VNIR with 768 bands
(2734 x 1312) x 768 x 2bytes = **5.1GB image**

1 sample every 5 min
Twilight to twilight on June 21 = **1TB**

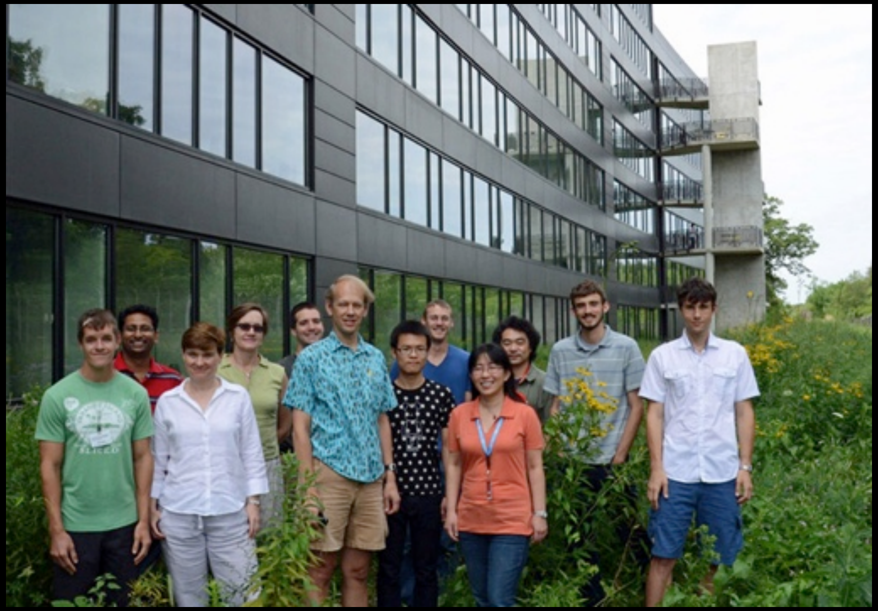
We need a parallel computer with each sensor!

Argonne NATIONAL LABORATORY

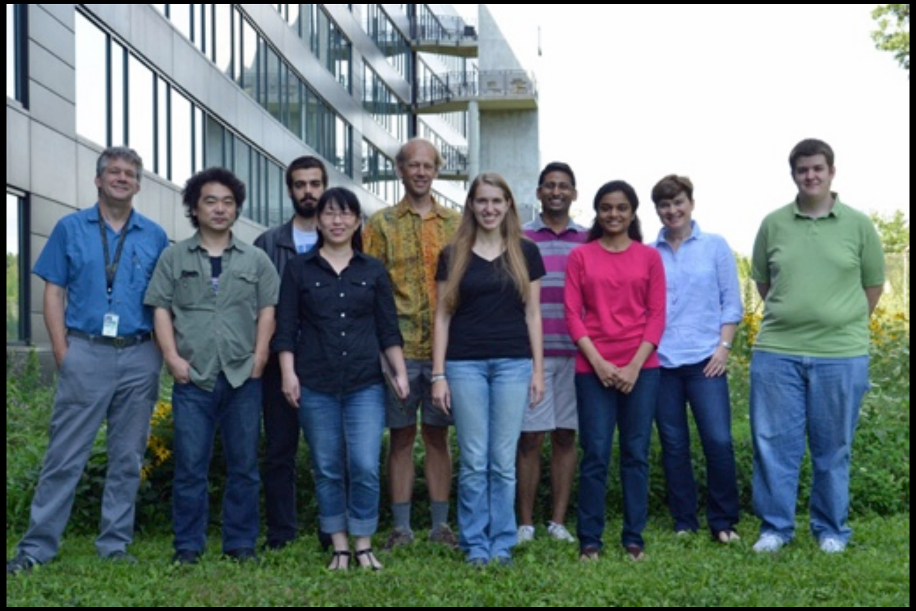
“Out on the edge you see all kinds of things you can't see from the center. [...] Big, undreamed-of things - the people on the edge see them first.”

- Kurt Vonnegut, *Player Piano*

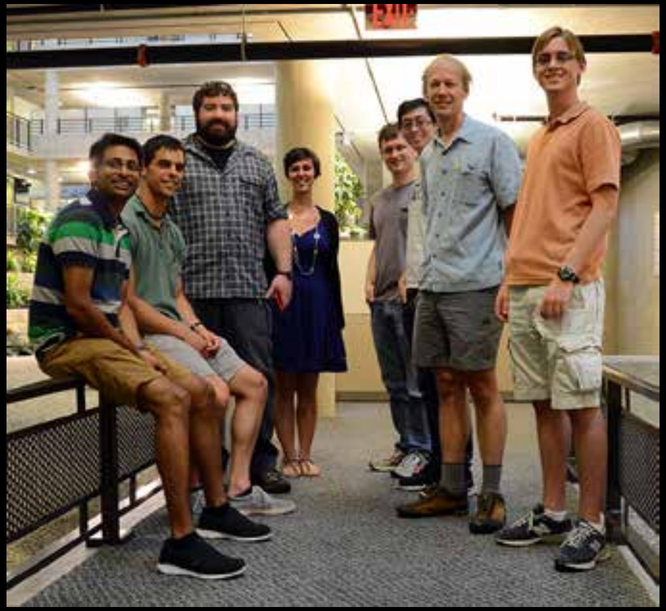
Argonne NATIONAL LABORATORY



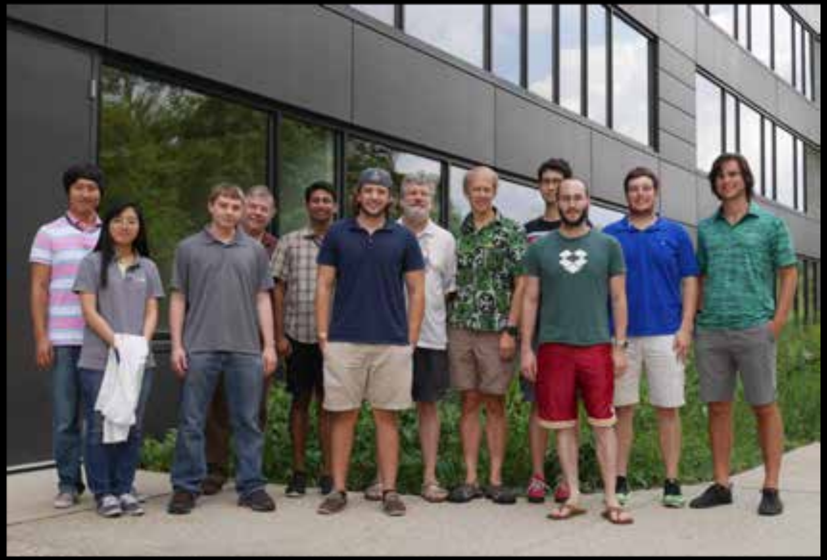
2013



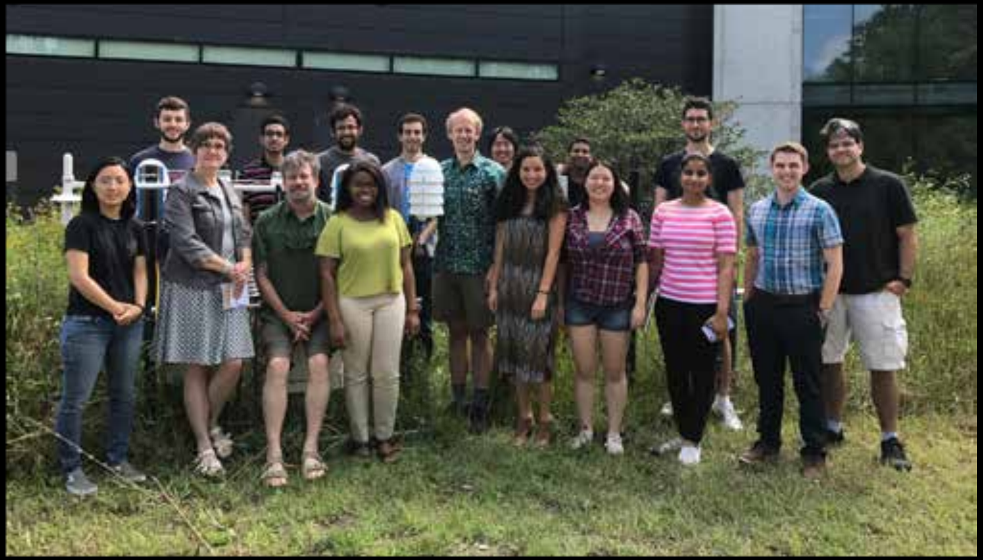
2014



2015



2016



2017

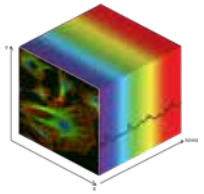
The Edge + Machine Learning A Revolution

Sensors

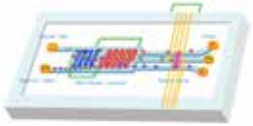
LIDAR

Software Defined Radios

Cameras



Hyperspectral Imaging



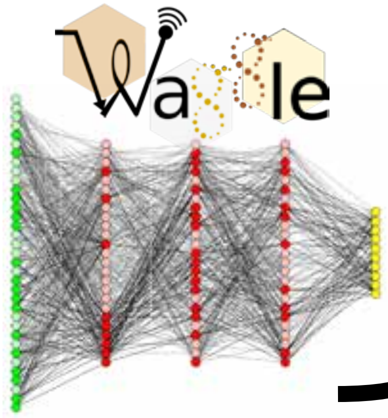
Microfluidic Sensors

Actuators

Servos

Dynamic adaptation

Powerful
Parallel Edge Computing



Artificial Intelligence
Deep Learning Inference

Semantic Output

Reduced, Compressed data

New inference (program code)

Edge computing and deep learning with feedback for continuous improvement

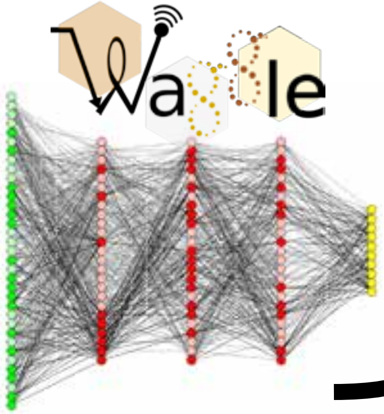
HPC/Cloud

Deep Learning Training

The Edge + Machine Learning A Revolution

Facility

Powerful
Parallel Edge
Computing



Semantic
Output

Artificial Intelligence
Deep Learning Inference

Edge computing and deep learning with feedback for continuous improvement

Reduced, Compressed data

New inference (program code)

HPC/Cloud

Deep Learning
Training

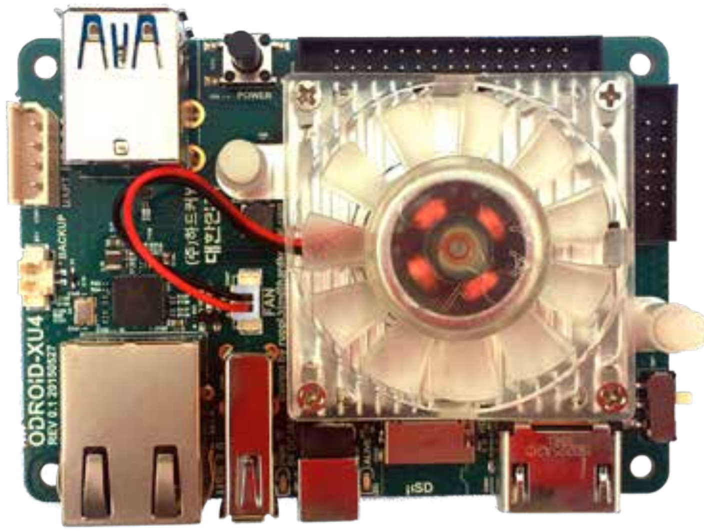
Actuators

Servos

Dynamic
adaptation

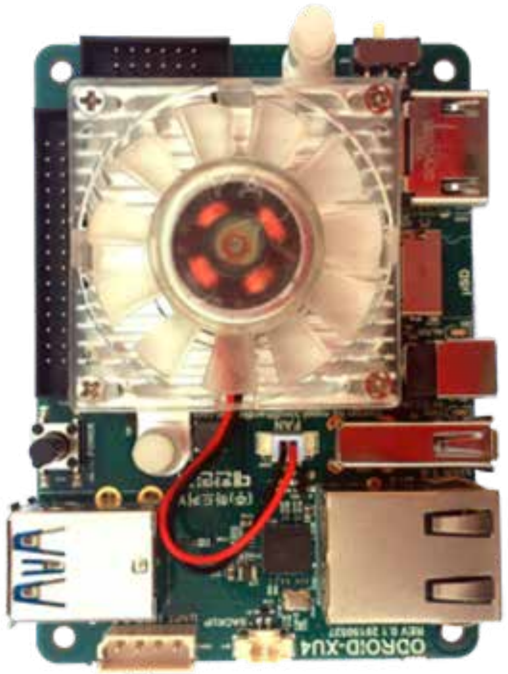
Pete Beckman beckman@anl.gov

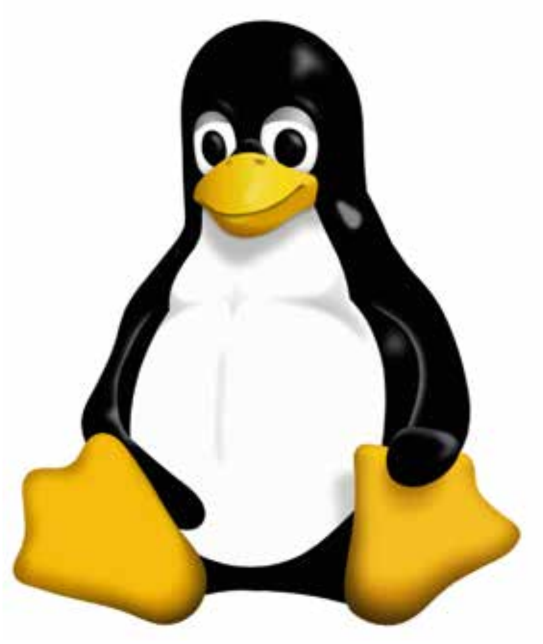
Why Live on the Edge?



- **More data than bandwidth**
  - Spallation neutron source, light source, HD Cameras, LIDAR, radar, hyperspectral imaging, grid micro-synchrophasors, etc.
- **Latency is important**
  - Quick local decision & actuation; adaptive sensing & control systems
- **Privacy/Security requires short-lived data: process and discard**
  - Compromised devices have no sensitive data to be revealed
- **Resilience requires distributed processing, analysis, and control**
  - Predictable service degradation, autonomy requires local (resilient) decision
- **Quiet observation and energy efficiency**
  - Vigilant sensors, transmit only essential observations, not big data streams

When a Computer + Linux is Not Enough…

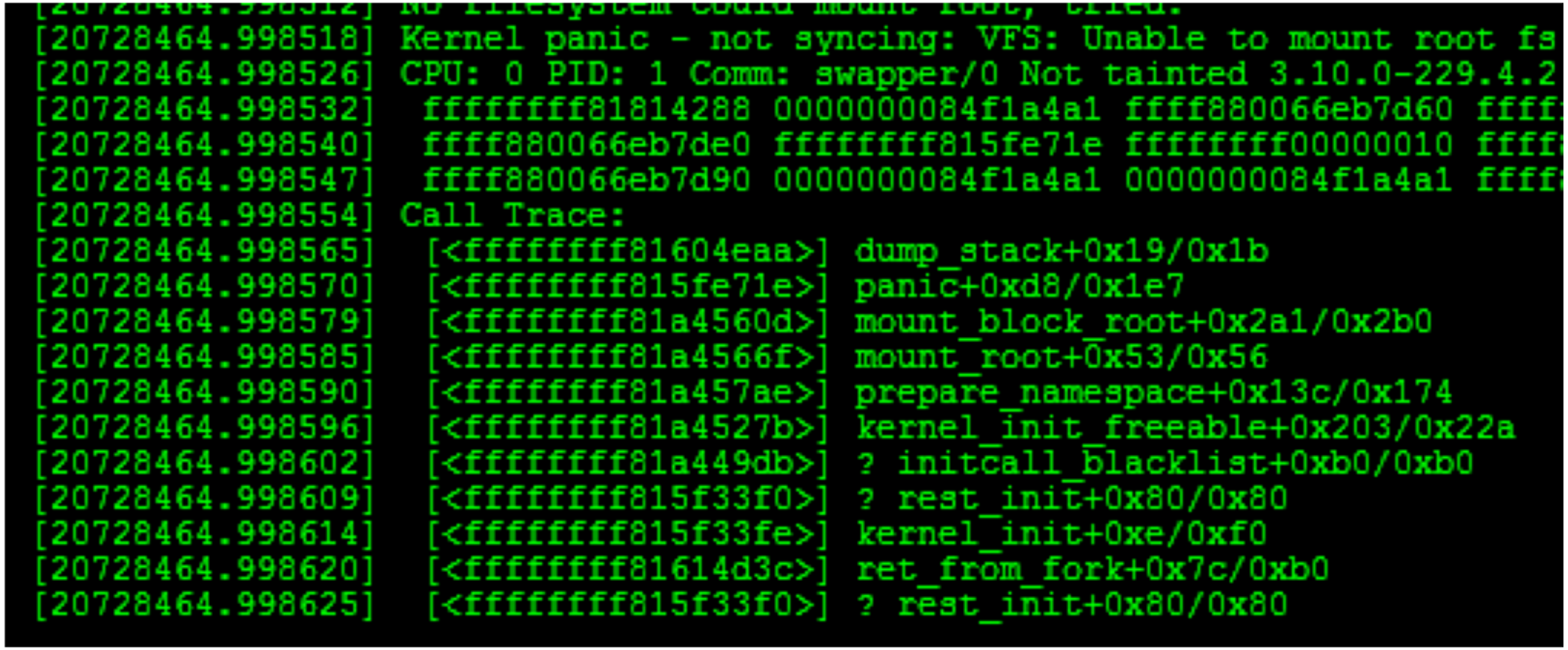




Challenging Design Contradiction

$\perp$

- Experimental ML/GPU software fails often
- Edge Devices are remotely deployed



Smoothing Out The Rough Edges

A Pocket-Sized Controller for Edge Computing

The Waggle Manager (WagMan)

- Borrowed BG/Q control system ideas
- Designed mini “rack controller”
  - Devices can be disconnected
  - Devices can be power cycled
- “Deep Space Probe” design
  - Heart beat signals to each device
  - Alternative boot image / safe mode
  - Current and voltage monitoring
  - Environmental monitoring
- Strict cybersecurity design

Waggle: Argonne's *Edge Computing* Platform

Bring Parallel Computing to the Edge

- Supports powerful, parallel computation at the edge
  - Computer vision and deep learning frameworks (Caffé, TensorFlow, OpenCV)
  - Supports edge-optimized & experimental computing
    - ML hardware, GPUs, neuromorphic, FPGAs, etc.
- Open Source, open interfaces
- Integrating advanced sensors easy, with plug-in architecture
- Robust remote system management subsystem
- Manufactured at local electronics company

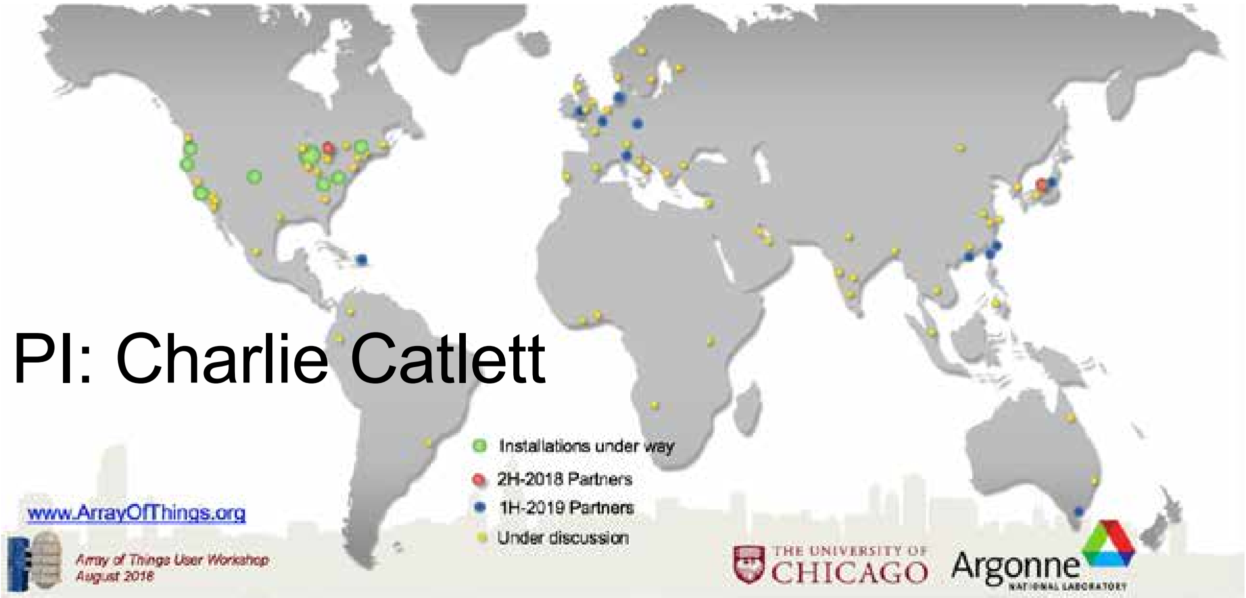
- ~5 years of development by team at Argonne National Laboratory


LEMONADE
Ice Cold
$1.00
per glass
#FOODFORALL
LEMONADE
for sale
$1=2 meals!
#FOODFORALL
$1=2 MEA

The *Array of Things* Project is Deploying Hundreds of Waggle-based Nodes in Cities

UChicago / National Science Foundation

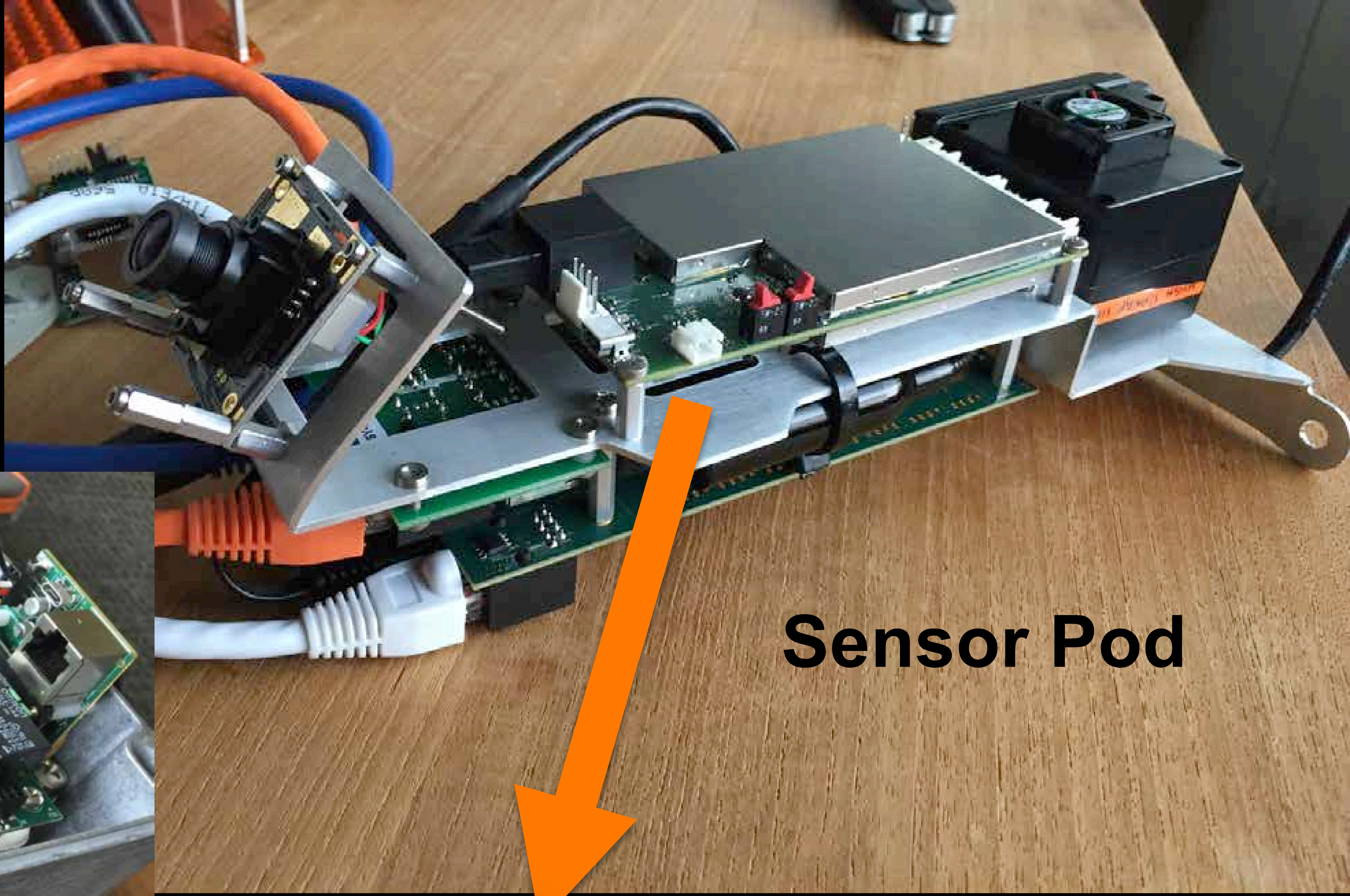
- 500 nodes will be deployed in Chicago
- Pilot Cities: Denver (Panasonic), Seattle, Portland, Palo Alto, Detroit, Syracuse, Tokyo, Chapel Hill.
- 20+ other cities preparing for pilot projects
- Nodes have 2 cameras, one up, one down
- An instrument to understand urban issues



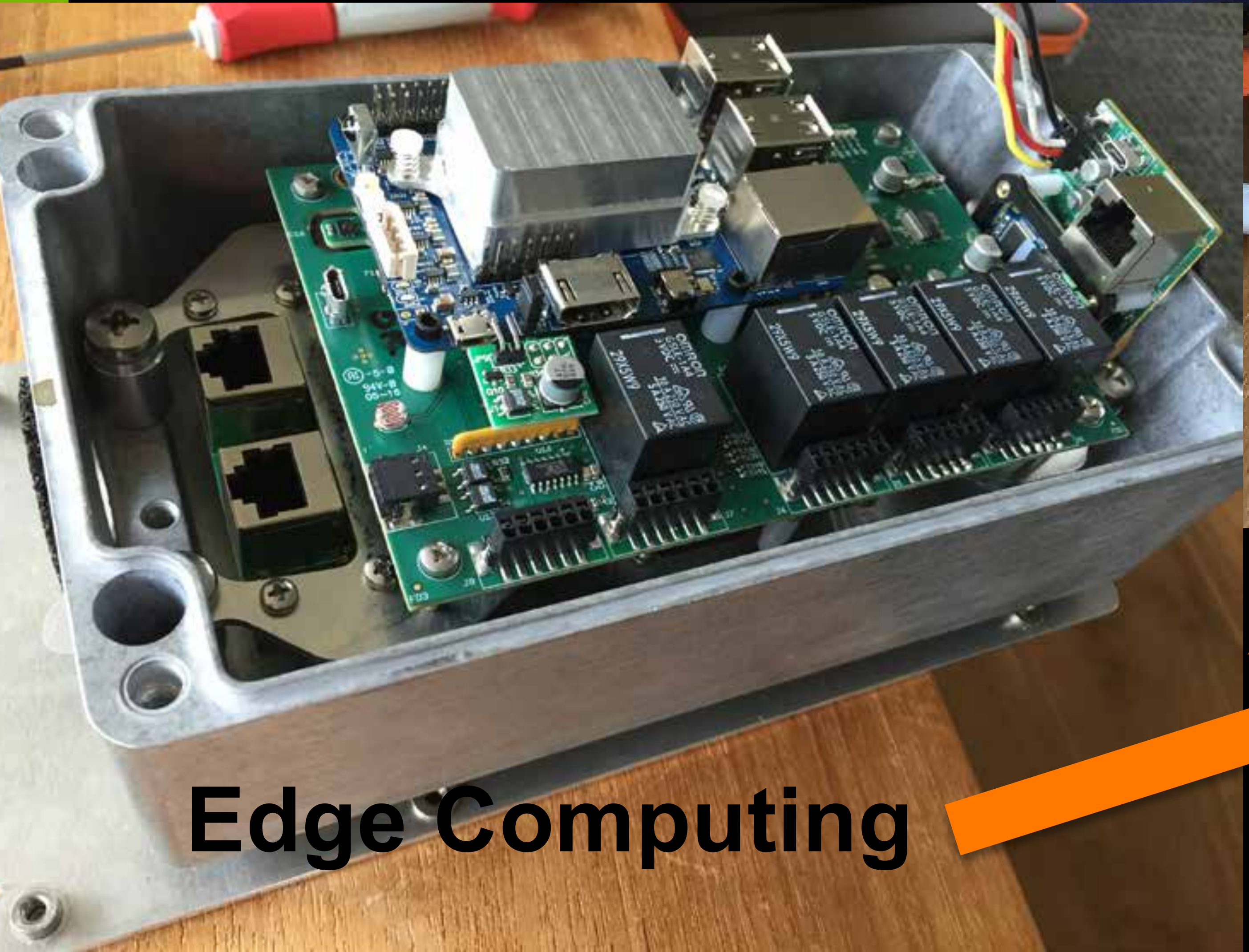
PI: Charlie Catlett

Rack of AoT nodes headed to Detroit

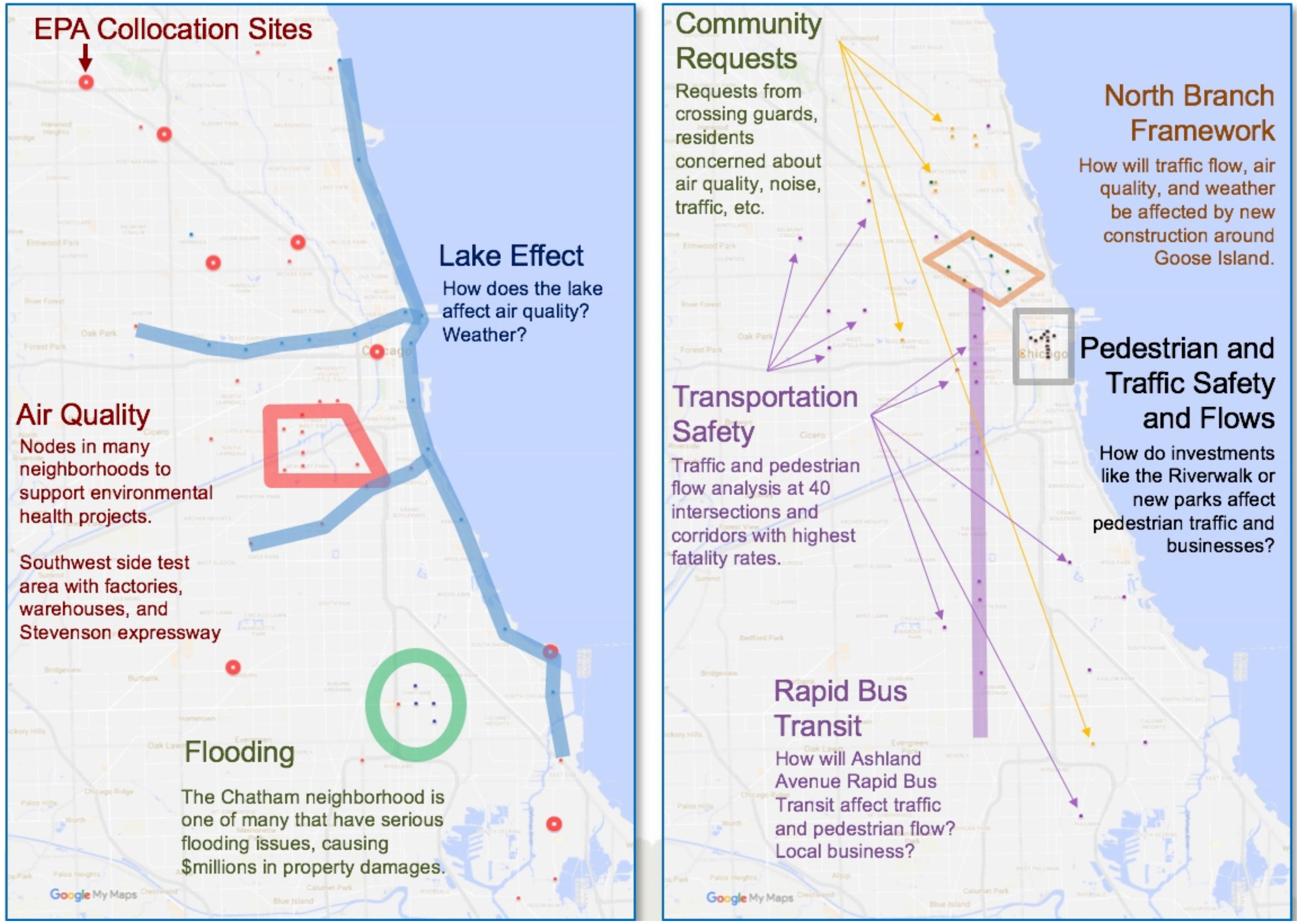
Array of Things Teardown



Sensor Pod

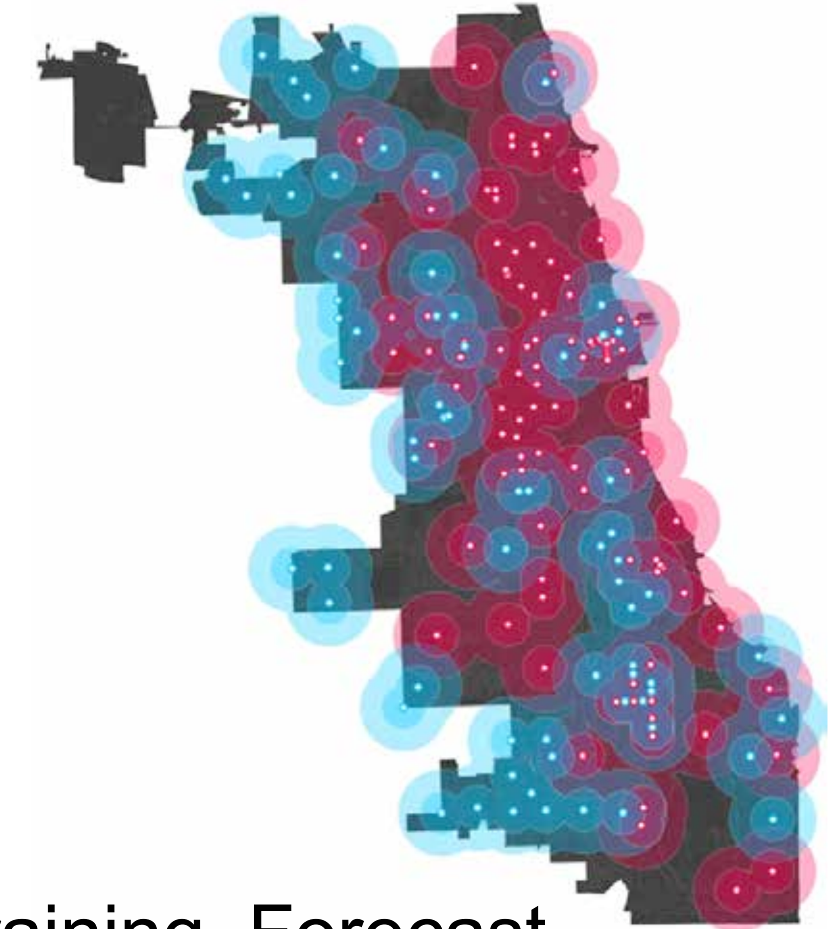


Edge Computing



Initial 105 AoT node locations, showing that locations are selected in groups as part of specific science investigations

Current (red) and 60 of 100 additional planned (blue) AoT nodes. Both 1km and 2km buffers are shown, illustrating that even with 200 nodes over 95% of Chicago's residents will live within 2km of a node and over 75% will live within 1km.



- **HPC**: Training, Forecast, Optimization, Observation
- **Edge**: Inference, Actuation, Lightweight Learning

GIS map created by A. Laha, Center for Spatial Data Science, University of Chicago

The Computing Continuum

IoT/Edge — Fog — HPC/Cloud

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
Memory	0.5K	256K	8GB	32GB	256G	32TB	16PB
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

Transportation

Advanced computer vision to understand pedestrian movement, eventually to predict dangerous interactions with vehicles

Research Credits:
Zeeshan Nadir (Purdue PhD Student @ ANL, 2017)
Nicola Ferrier (ANL Scientist)

Mask r-CNN [He, 2017]
trained on COCO dataset

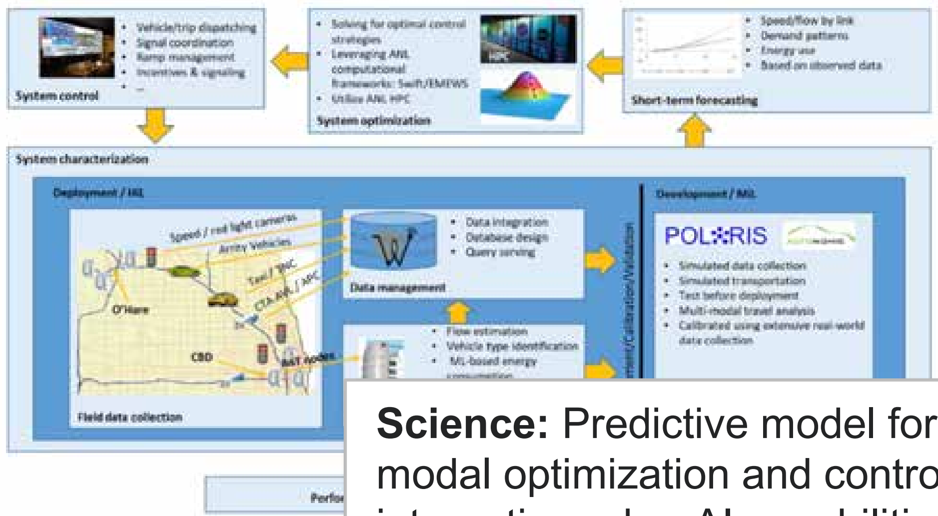
Research Credits:
Yongho Kim, Seongha Park (Purdue PhD Students @ ANL, 2018)
Pete Beckman, Nicola Ferrier (ANL Scientists)

AoT Image from Lake Shore Drive

Transportation

Science: Prototype model of at-grade crossing with impact analysis (interrupt duration and impact; emergency vehicles delayed)

Objective: Prioritize among hundreds of at-grade crossings in context of $1B planned investments to improve rail throughput by eliminating key at-grade crossings.



Science: Predictive model for multi-modal optimization and control, integrating edge-AI capabilities with traditional transportation data, coupled with HPC models and control systems.

Deployment: Integrate transportation measurements from AoT/Waggle (density, flow, vehicle mix, parking) with live traffic data and traffic model around O'Hare International Airport.

$3.2m

Funding: Illinois DOT

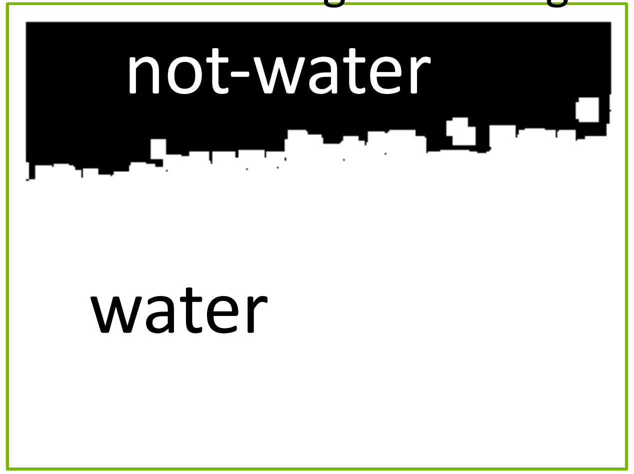
Partners: Argonne, UChicago, Chicago Metropolitan Agency for Planning, Chicago DOT

Funding: EERE VTO

Partners: Argonne, Chicago DOT, Chicago Dept. of Aviation, Chicago Dept. of Innovation and Technology, Arity

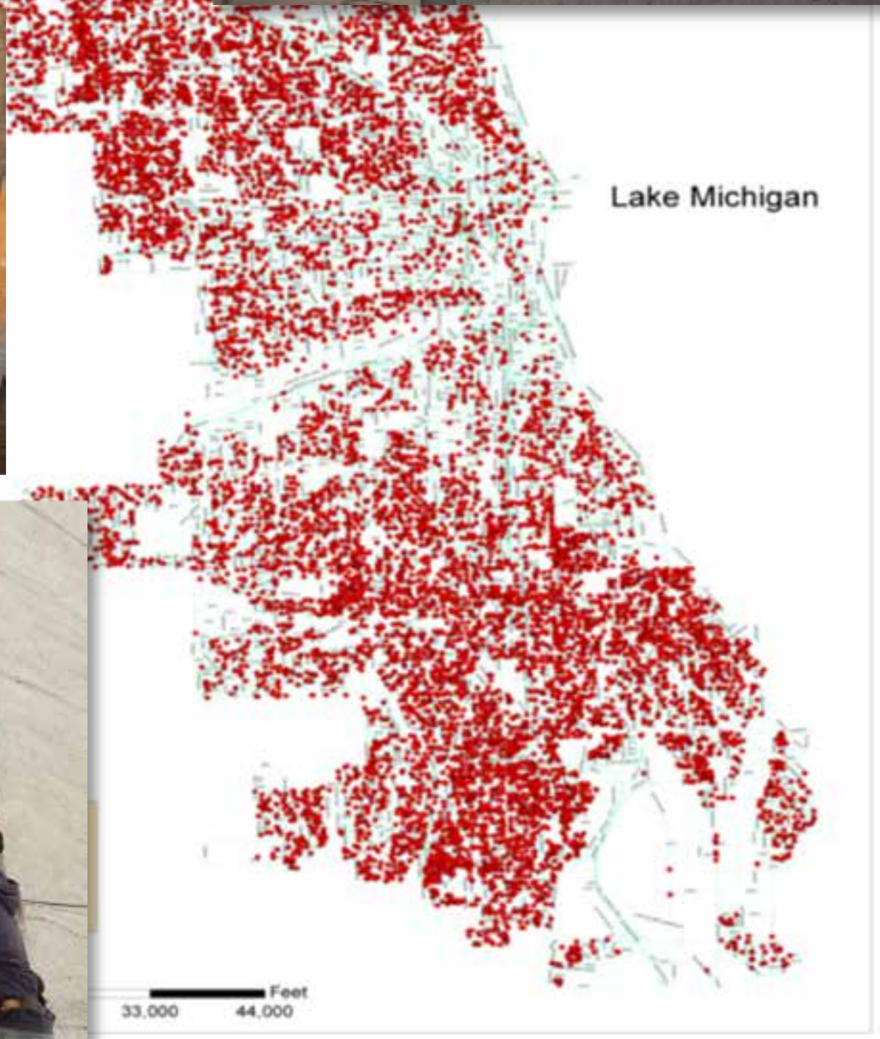
Hydrology: Flooding

50 consecutive frames to flood water and segment image



Using advanced computer vision to detect surface flooding

Research Credits:
Ethan Trokie (Northwestern Undergrad Student @ ANL, 2017)
Vivien Rivera (Northwestern PhD Student, SCGSR 2018)
Nicola Ferrier (ANL Scientist)
Rajesh Sankaran (ANL Scientist)



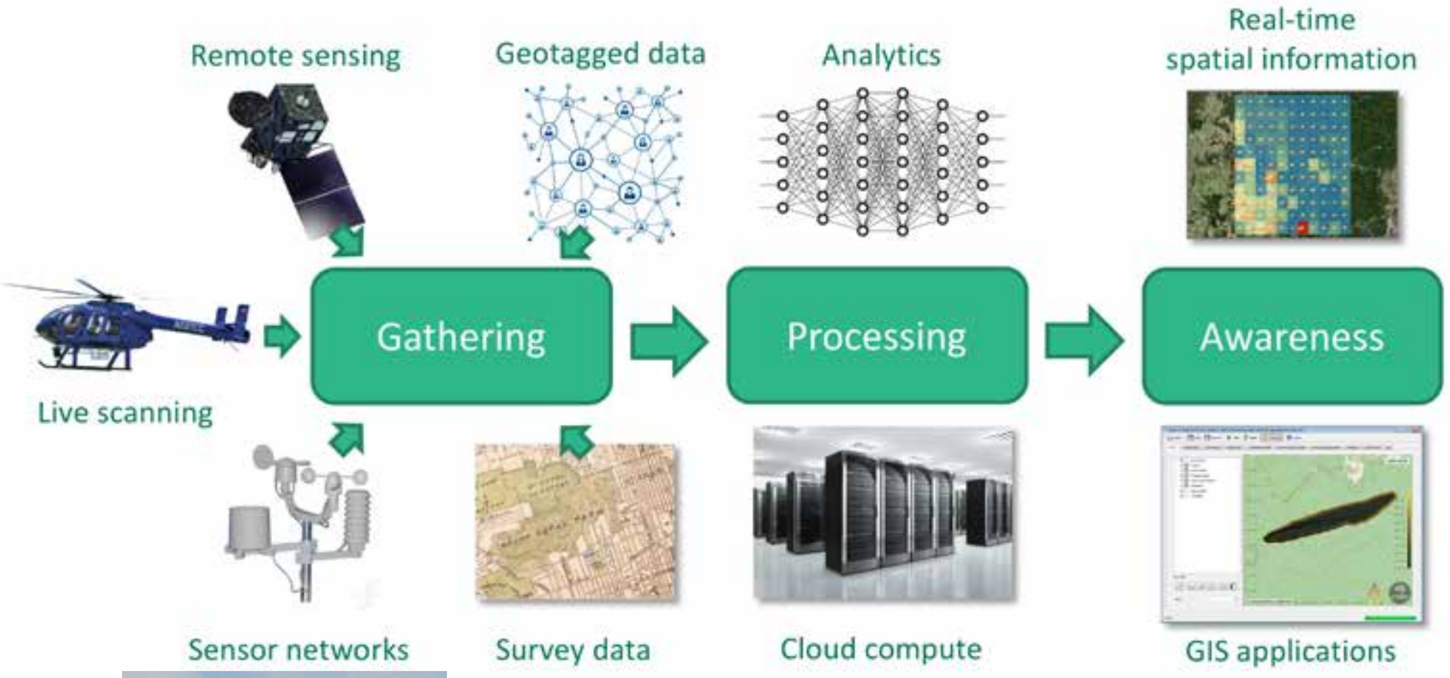
Live HPC Flood Modeling and Prediction?

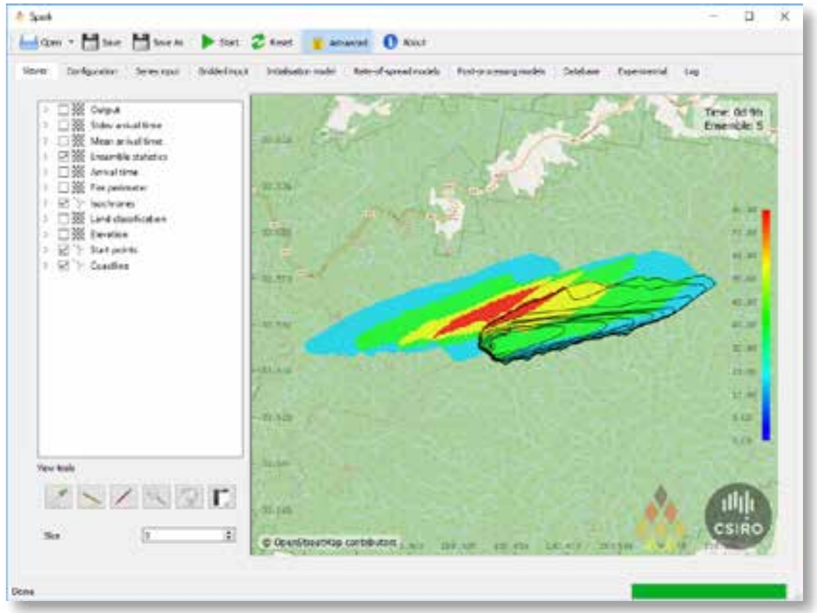
Work with Aaron Packman and William Miller (Northwestern University)
Cristina Negri, Rajesh Sankaran, Nicola Ferrier (Argonne)

Waggle Nodes in Tuley Park, Chicago

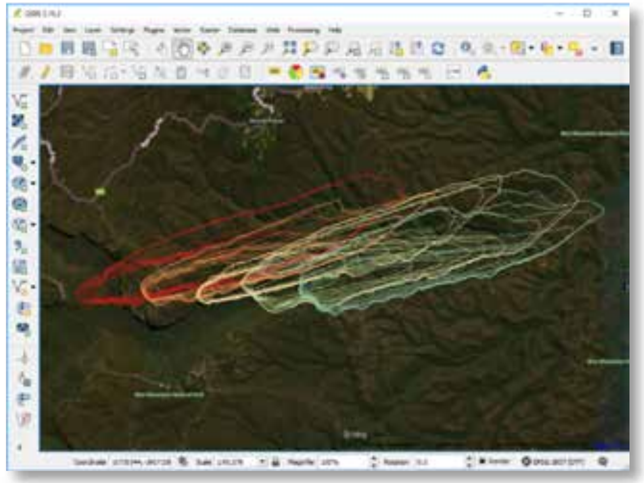
Disaster: Flood Fire

Partnership with CSIRO Australia (MOU, visiting postdoc)

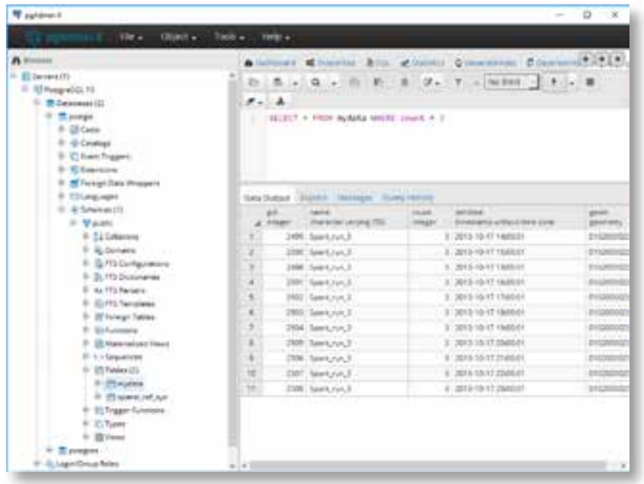




Impact analysis from ensemble wildfire simulations



GIS visualisation



Integration with PostGIS database

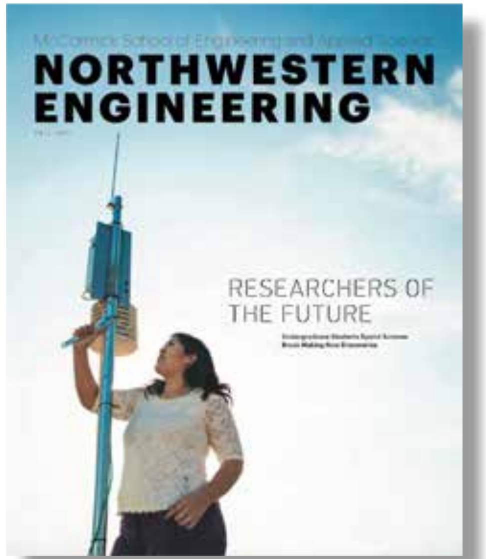
Nikhil Garg

Live HPC Modeling and Prediction…

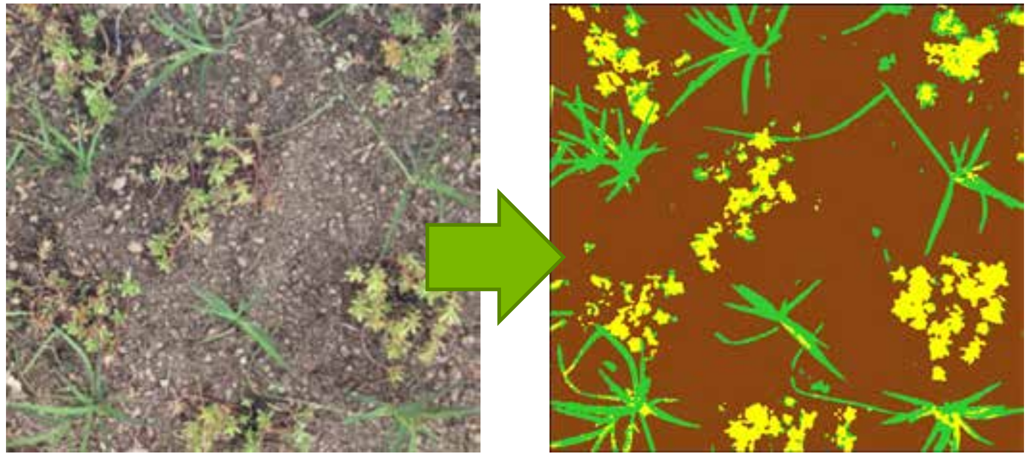
- Basis of D61 natural hazard applications:
  - Wildfire and wildfire impact (Spark)
  - Flood and coastal inundation (Swift)

Earth Modeling: Ecosystem Response

Advanced sensors and computer vision to monitor pristine prairie



Research Credits:
Vivien Rivera (Northwestern Univ. PhD Student, 2018)
Aaron Packman, Bill Miller (Northwestern Univ. Professors)
Pete Beckman (ANL Scientist)



Original Auto Segmentation

Using advanced computer vision to monitor plants

Research Credits:
Renee Zha (Northwestern Undergrad Student @ ANL, 2017)
Zeeshan Nadir (Purdue PhD Student @ ANL, 2017)
Nicola Ferrier (ANL Scientist)

Chicago Botanic Garden
Conservation Science Center

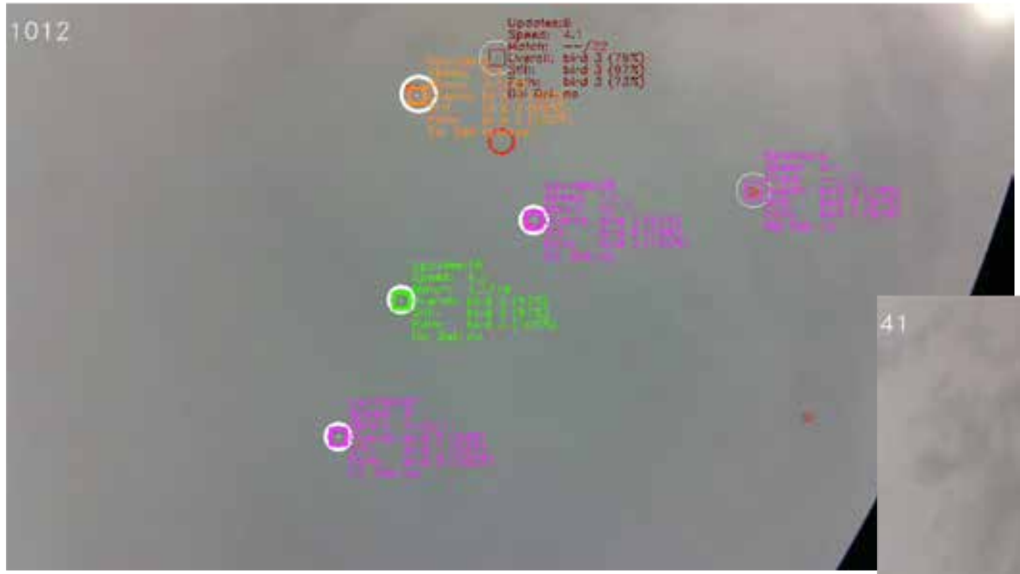
Undergrads Caeley and Jordan developed soil moisture sensor now deployed in Chicago

National Security

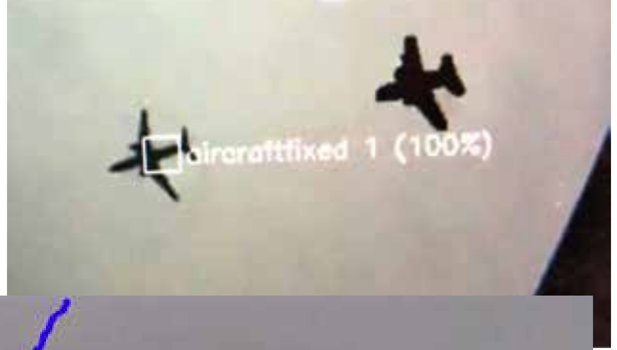
It's a Bird! It's a Plane! No… It's a Drone!

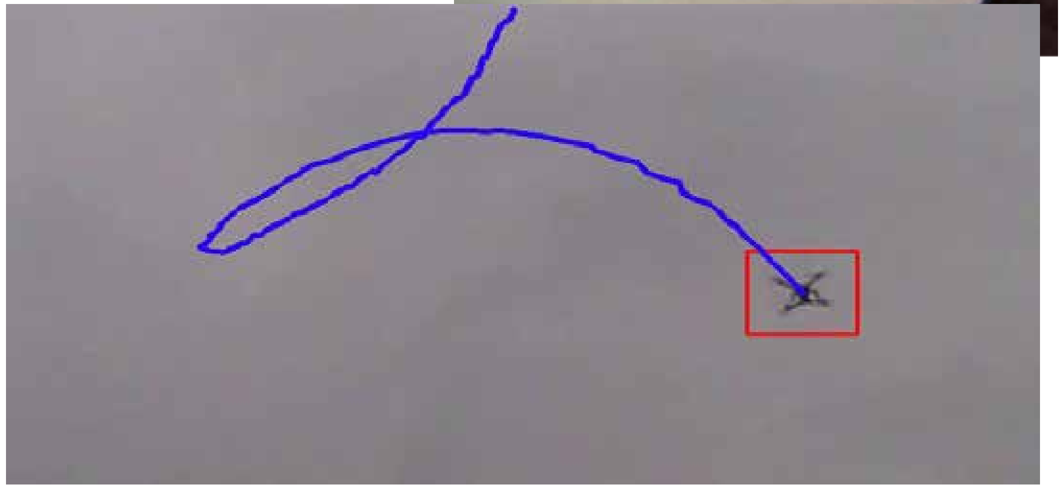
Advanced computer vision and machine learning to identify drones, birds, or fixed-wing aircraft.

Research Credits:
Sean Richardson (USAF visiting ANL)
Adam Szymanski (ANL Scientist)



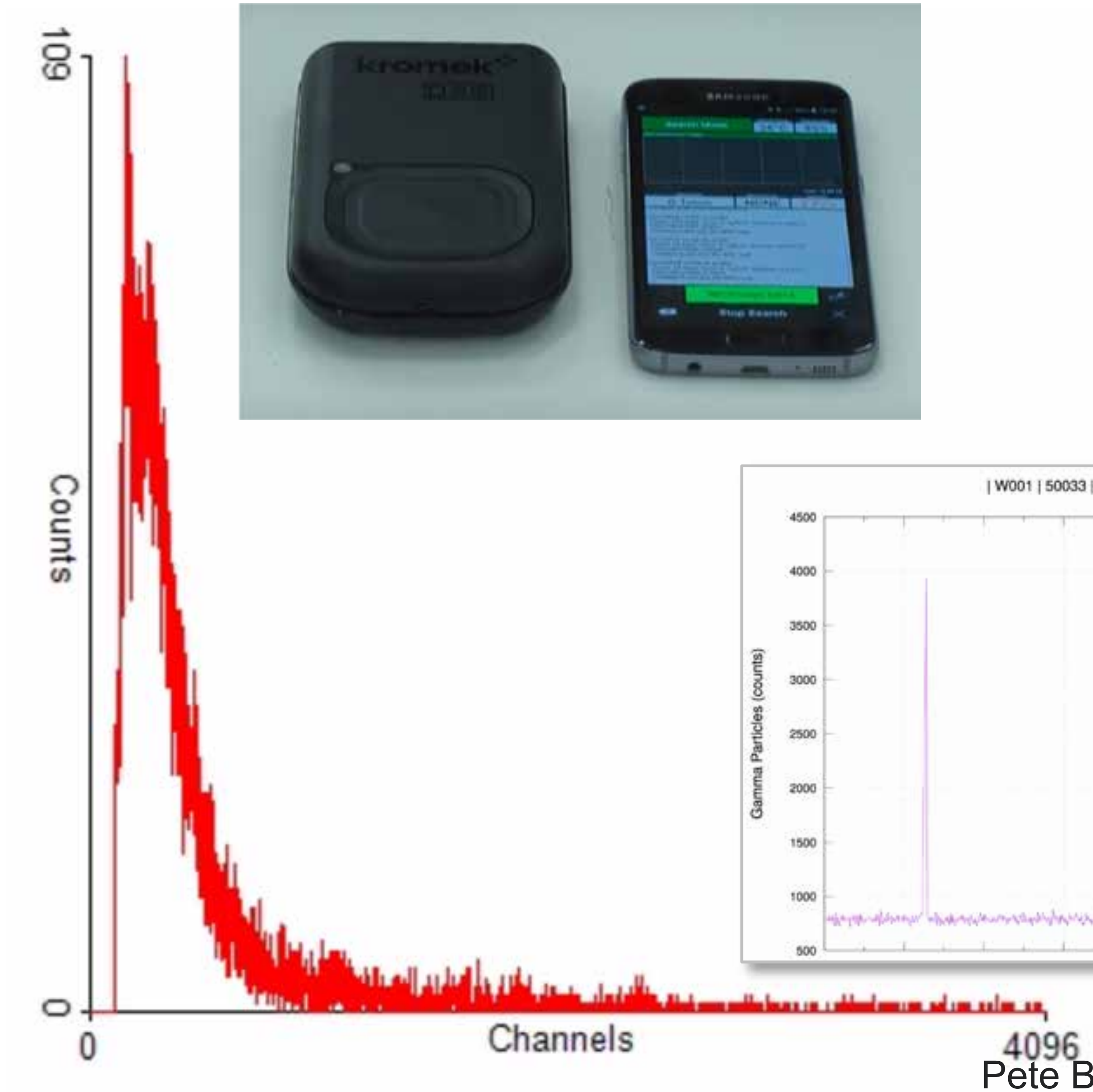
Object Classification

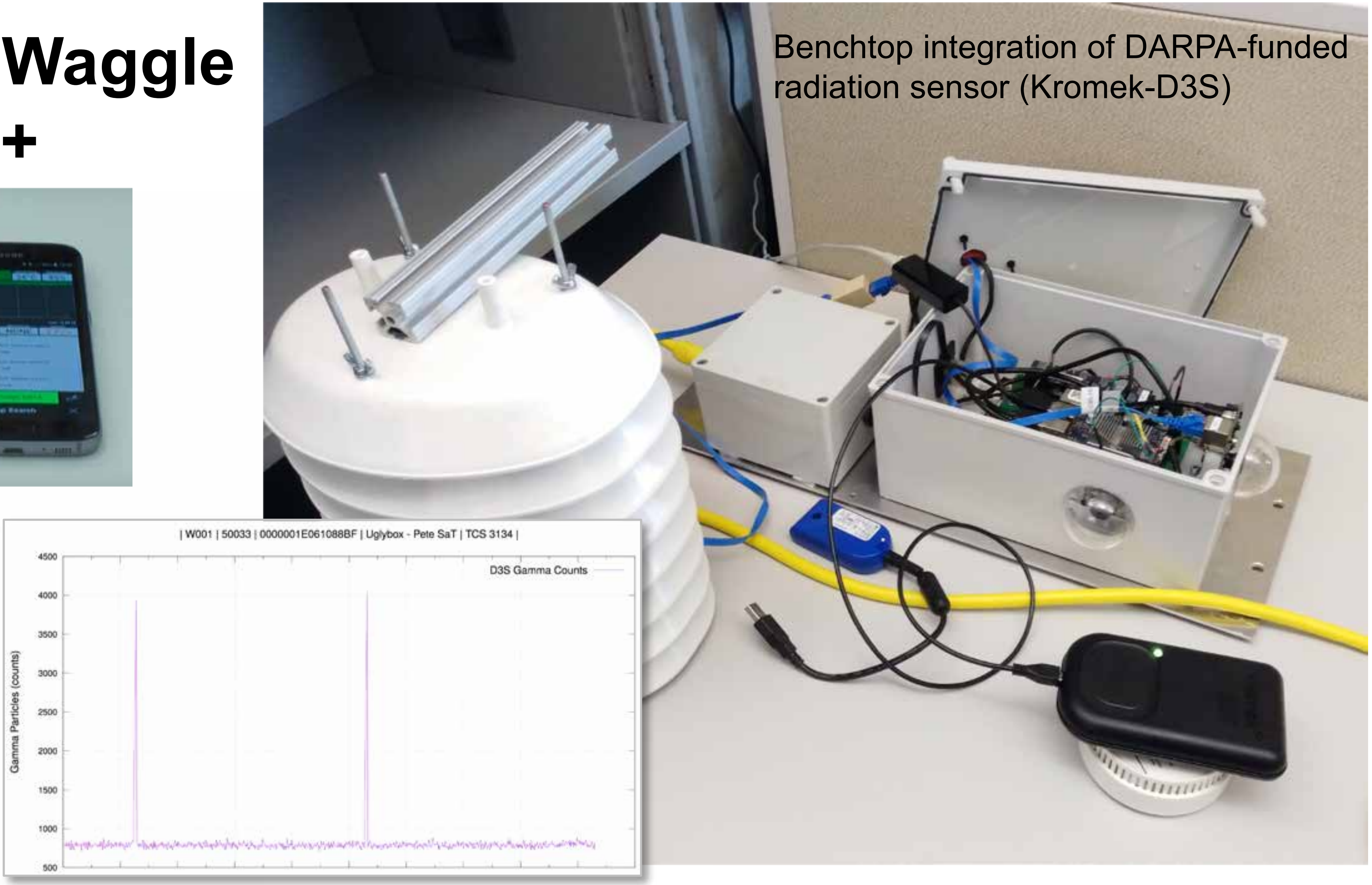




National Security

Testing Edge / Waggle DARPA SIGMA+





Benchtop integration of DARPA-funded radiation sensor (Kromek-D3S)

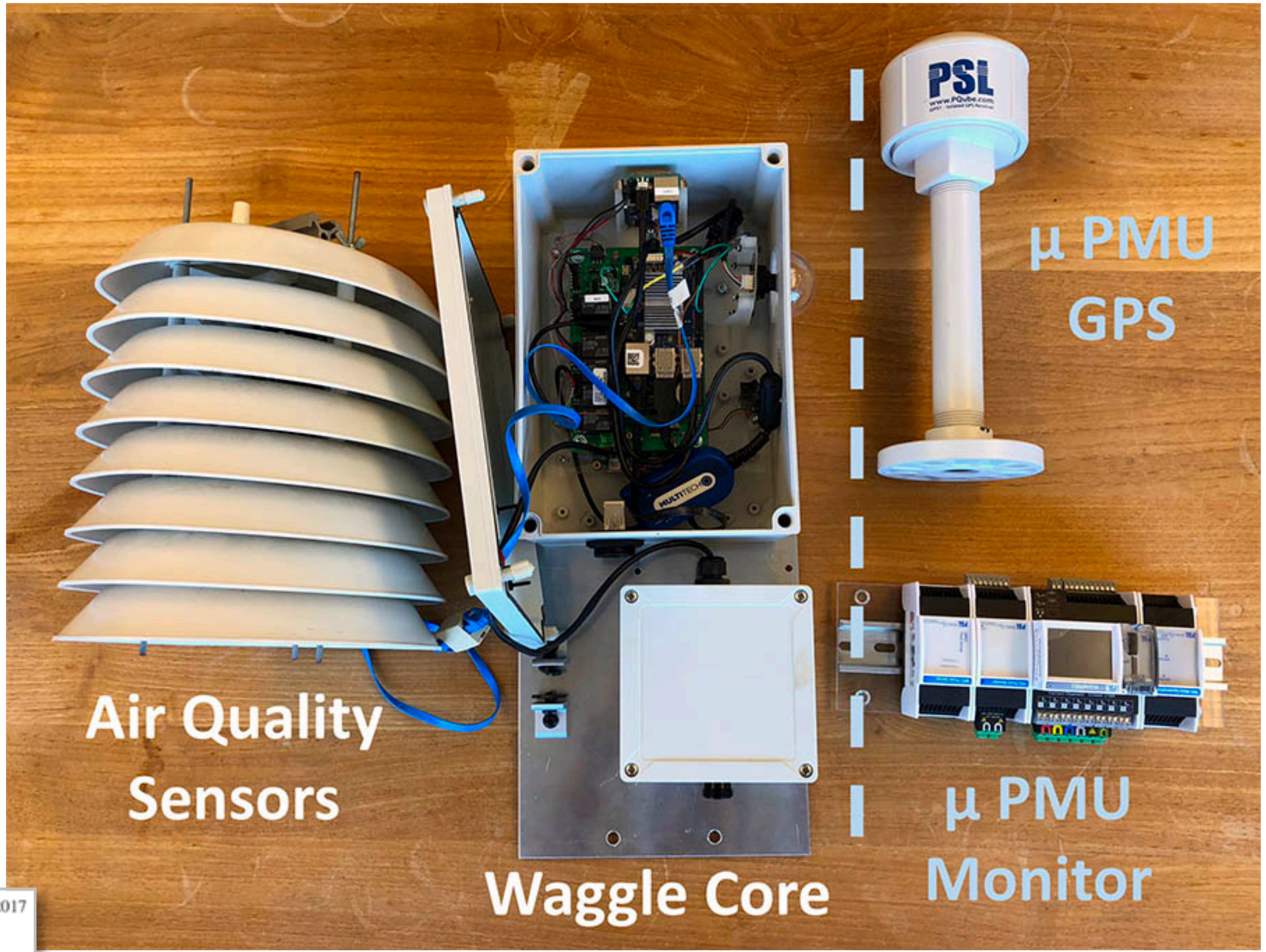
Energy: Power Grid

Exelon

CRADA Approved!!

Task Order 3 Approval
ASCR Concurrence Received 9/1/2018; OE Concurrence Received 9/11/2018;
SC-32 Concurrence Received 9/13/2018; Using Master CRADA without modification.

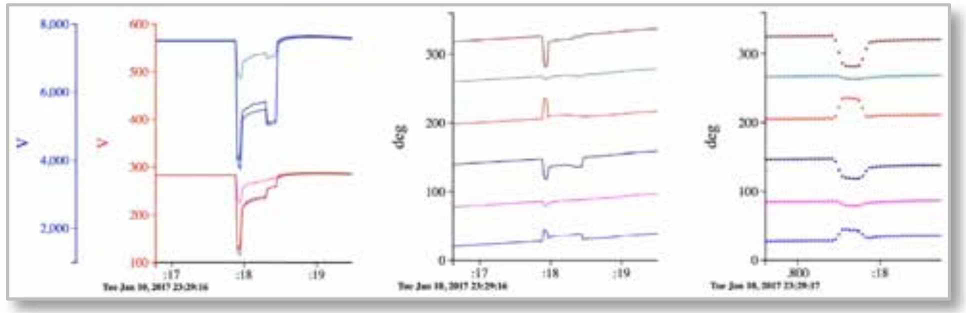
- **Load Forecasting**
- **Grid Stress**
- **Air Quality**



2926 IEEE TRANSACTIONS ON SMART GRID, VOL. 8, NO. 6, NOVEMBER 2017

Precision Micro-Synchrophasors for Distribution Systems: A Summary of Applications

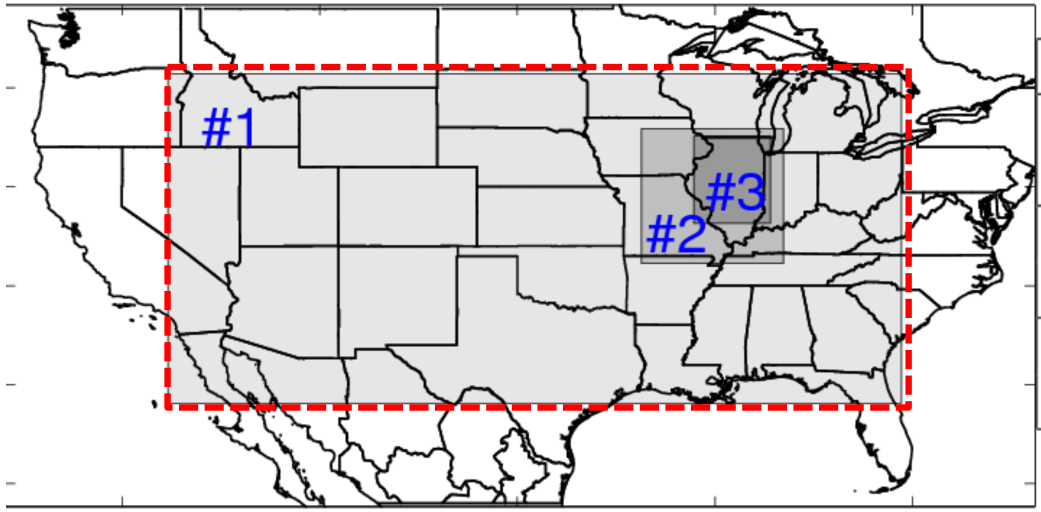
Alexandra von Meier, *Member, IEEE*, Emma Stewart, *Senior Member, IEEE*, Alex McEachern, *Fellow, IEEE*, Michael Andersen, *Member, IEEE*, and Laura Mehrmanesh



DOE
ARPA-E

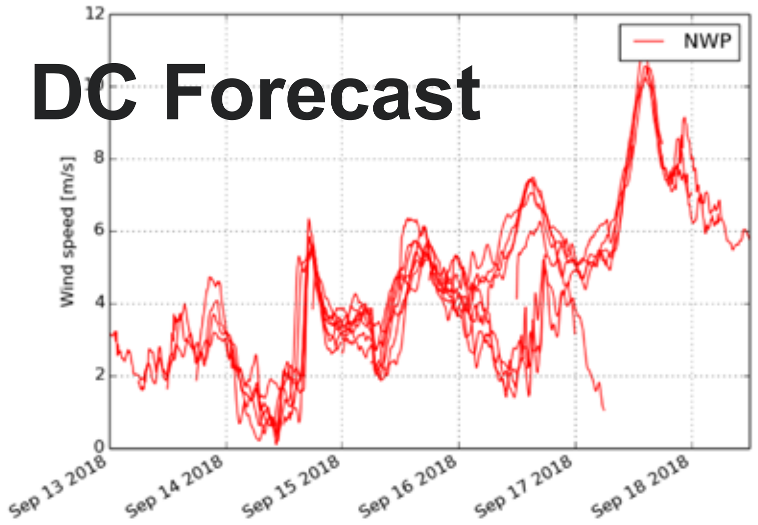
Computational Forecasting

Domain setup for HPC forecast



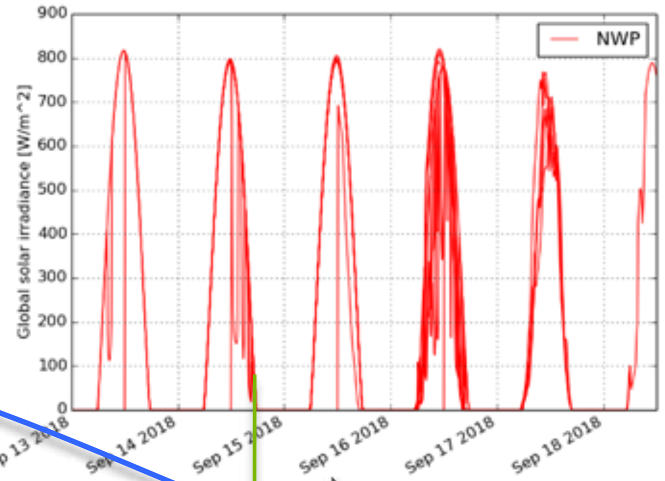
Grid/Size	Proc.	Walltime
#1 - 32 km^2 130 × 60	8	5362s
	16	3041s
#2 - 6 km^2 126 × 121	32	1599s
	64	1033s
#3 - 2 km^2 202 × 232	128	655s

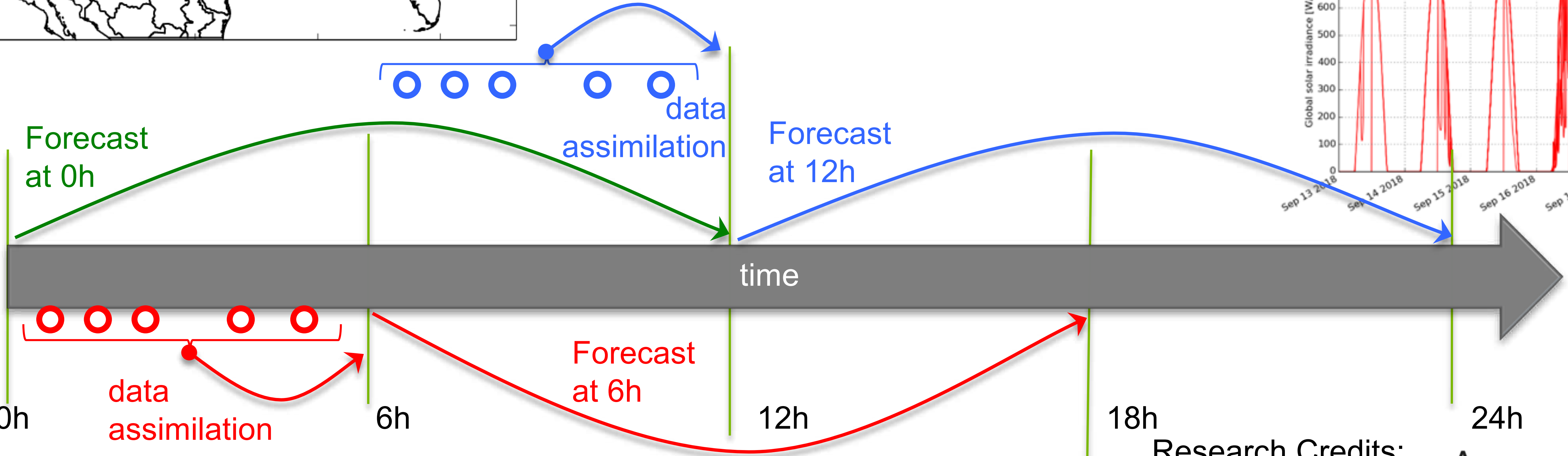
DC Forecast



Wind forecasts and measurements

Solar forecasts and measurements



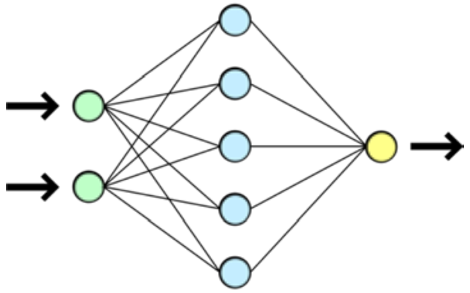


Research Credits:
Emil Constantinescu (ANL Scientist)

Earth Systems: Wind & Solar

Edge to HPC: Live & Historical Data for Forecasting

$$\overline{\mathbf{y}_*|\mathbf{X},\mathbf{X}_*,\mathbf{y}} = \mathbf{m}(X_*) + \mathbf{K}_{21}\left(\mathbf{K}_{11} + \quad\right)^{-1}(\mathbf{y} \quad \mathbf{m}(X))$$



GP & DNNs

Edge forecast

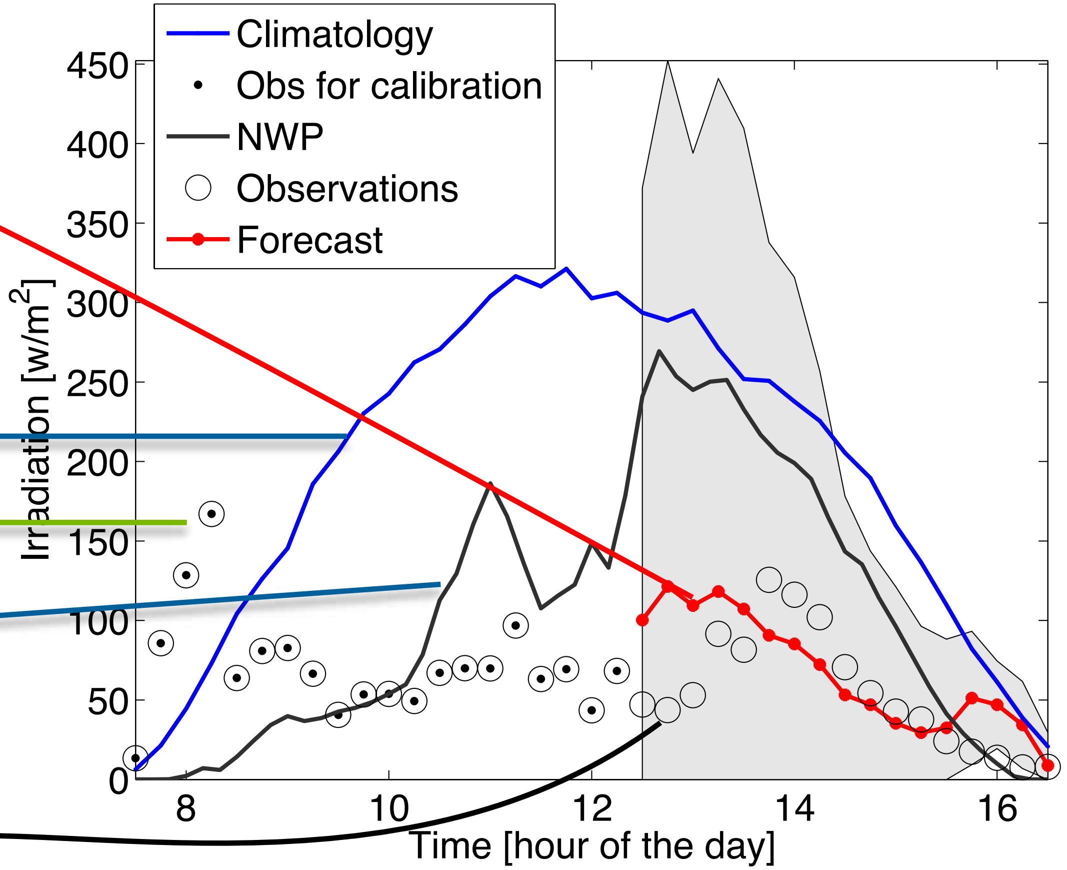
Numerical forecast

Historical data

Edge observations used in inference

Validation observations not used in inference

Conceptual Edge solar forecast



Research Credits:
Emil Constantinescu (ANL Scientist)

Manufacturing

Flame Spray Pyrolysis

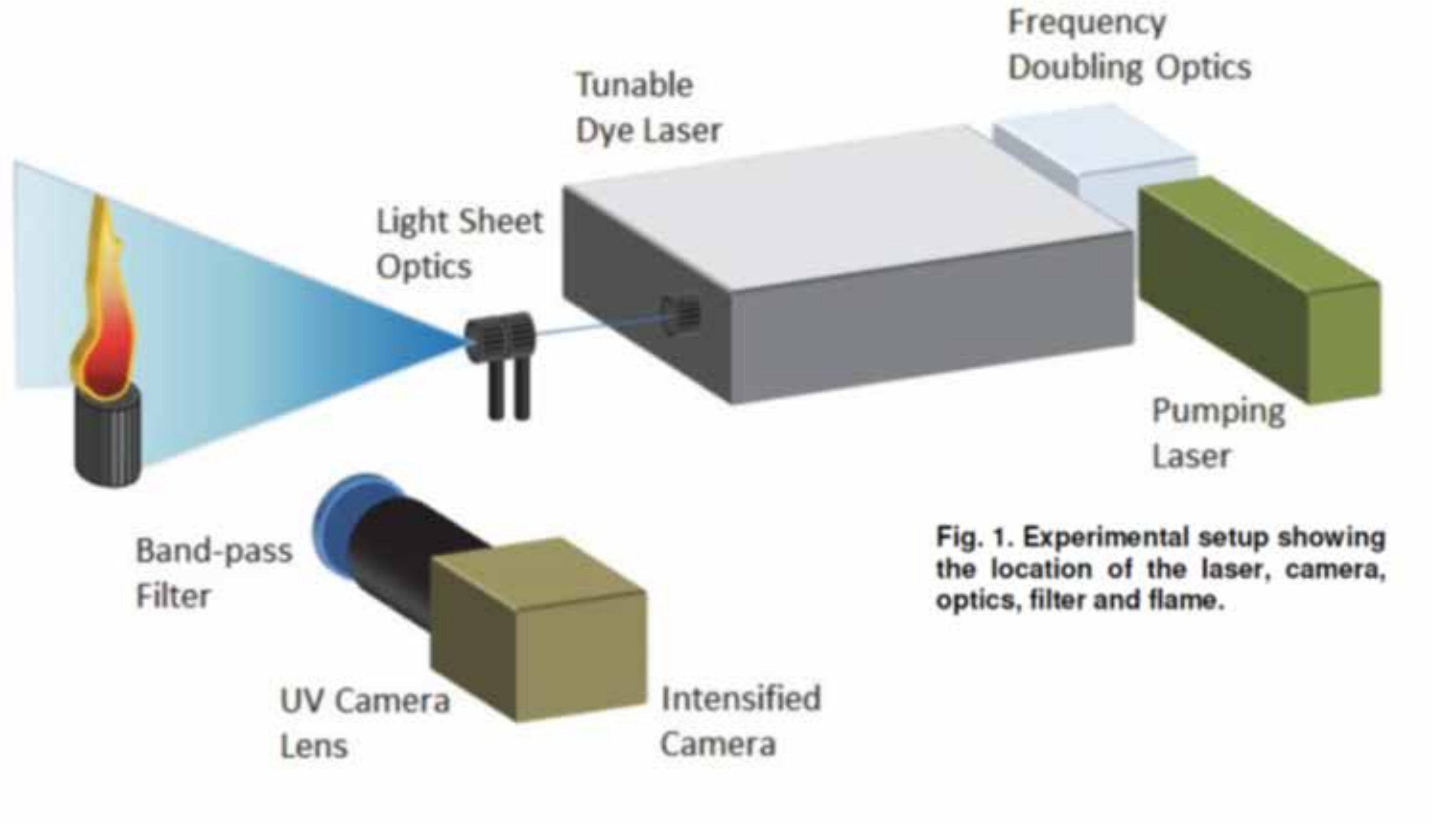
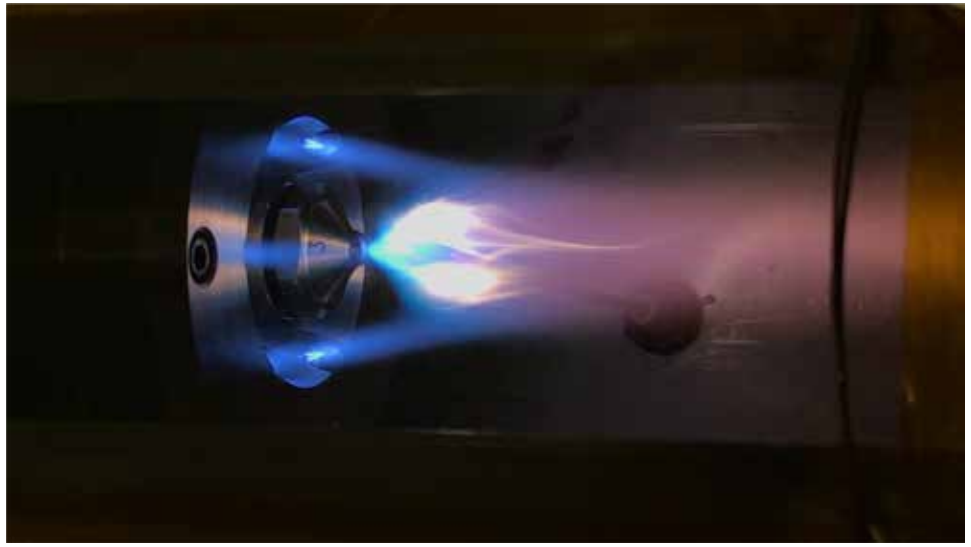
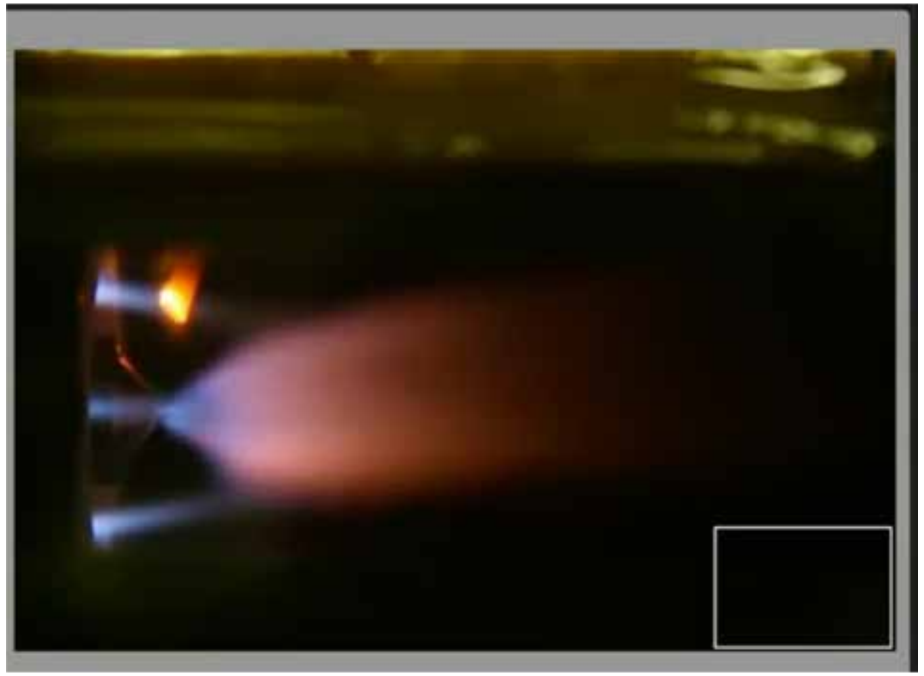


Fig. 1. Experimental setup showing the location of the laser, camera, optics, filter and flame.





LLZO transition

- Use data collected to date to develop ML/DL models
- Relate process parameters to output measures
- Optimize

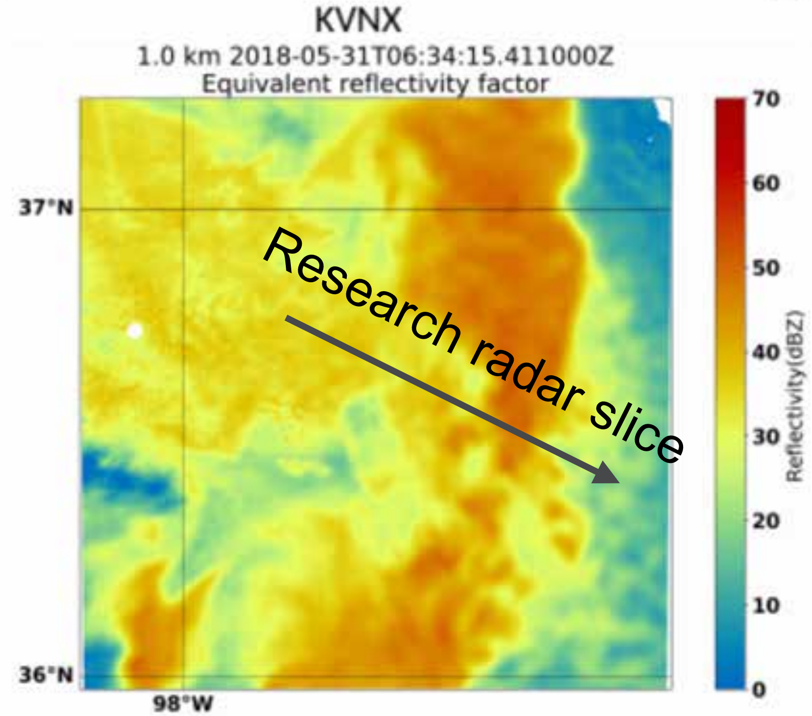
Research Credits:
N. Ferrier, J.Libera & S. Chaudhuri
Materials Engineering Research Facility, ANL

Atmospheric Science

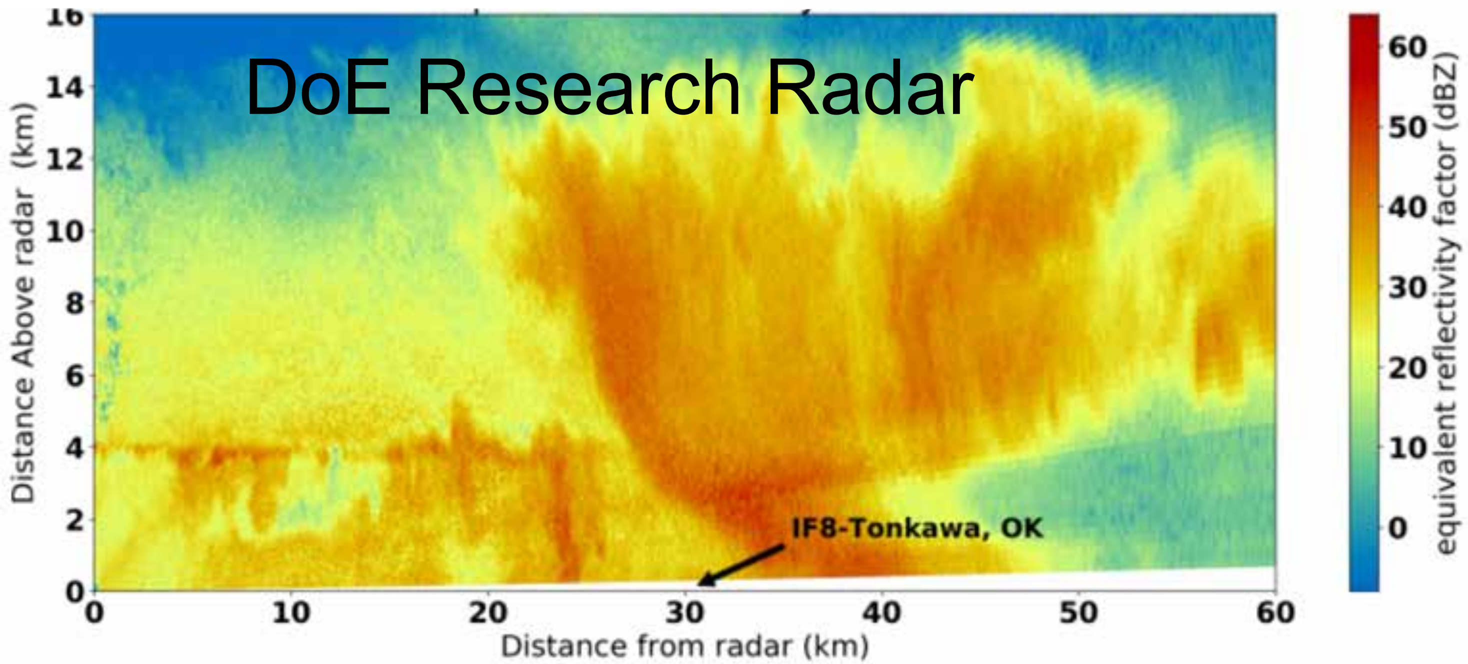
Adaptive sampling of the atmosphere

- Atmosphere sensing radars have a wide range of configurations.
- Ideal configuration depends on
  - Atmospheric scene:
    - hurricane, supercell, etc.
  - Phenomena of interest:
    - clouds, tornadoes, birds, bugs

Operational NOAA Radar



- AI@Edge needed to identify scene
- Automated slicing and dicing to reconstruct spatial structure using machine learning.



Research Credits:
Scott Collis, EVS Division, ANL

New NSF Funding

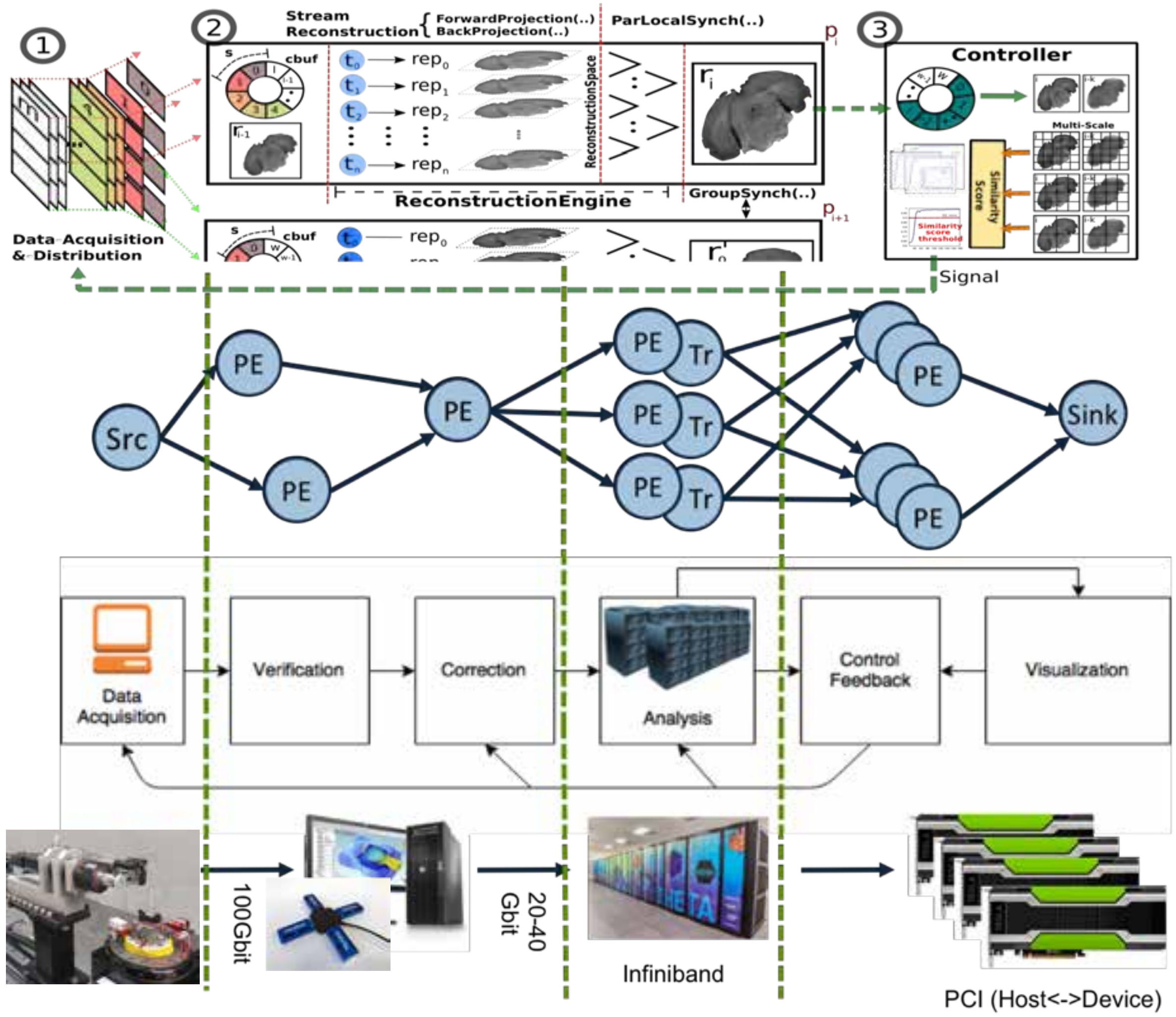
Facilities: Light Source

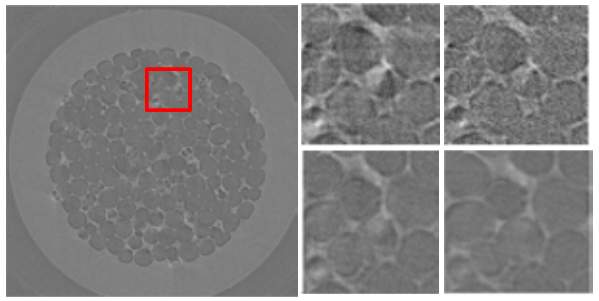
Data rates for APS-U will increase <u>several orders of magnitude</u>

Current Experiments:
Times vary: (from seconds to days/weeks)
- Many experimental parameters need to be optimizes
- Data analysis happens after experiment is finalized

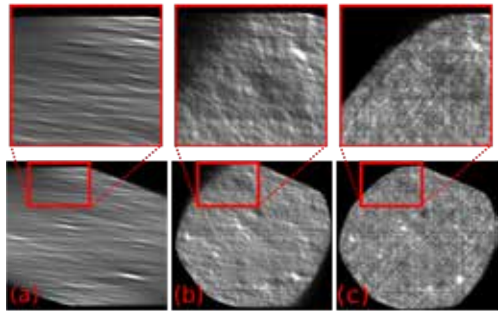
Data collection is in the dark
- Parameters are guessed (experience) and then optimized (repeated experiments)

Edge Computing and Experimental Steering
- Improving the science and the efficiency of the experiments
- Real-time data analysis and feedback, data verification, correction, normalization, and configuration parameter optimizations





Imaging aluminum foam (dynamic features) sample. Data acquisition and analysis parameters have significant affect on quality of reconstructed feature.



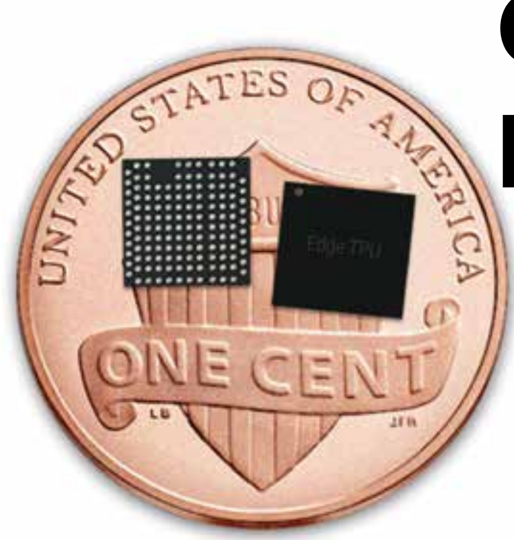
Real-time reconstruction of a shale sample. The scanning pattern and voxel coverage affect the reconstruction quality.

Research Credits:
Tekin Bicer, ANL

Investments in AI Hardware are Accelerating Change

At BOTH ends of continuum….

Missing: The programming framework for Edge-HPC Science



Google Edge TPU

July 2018

“Edge TPUs are designed to complement our Cloud TPU offering, so you can accelerate ML training in the cloud, then have lightning-fast ML inference at the edge. Your sensors become more than data collectors — they make local, real-time, intelligent decisions.”

cerebras

groq

FPGA

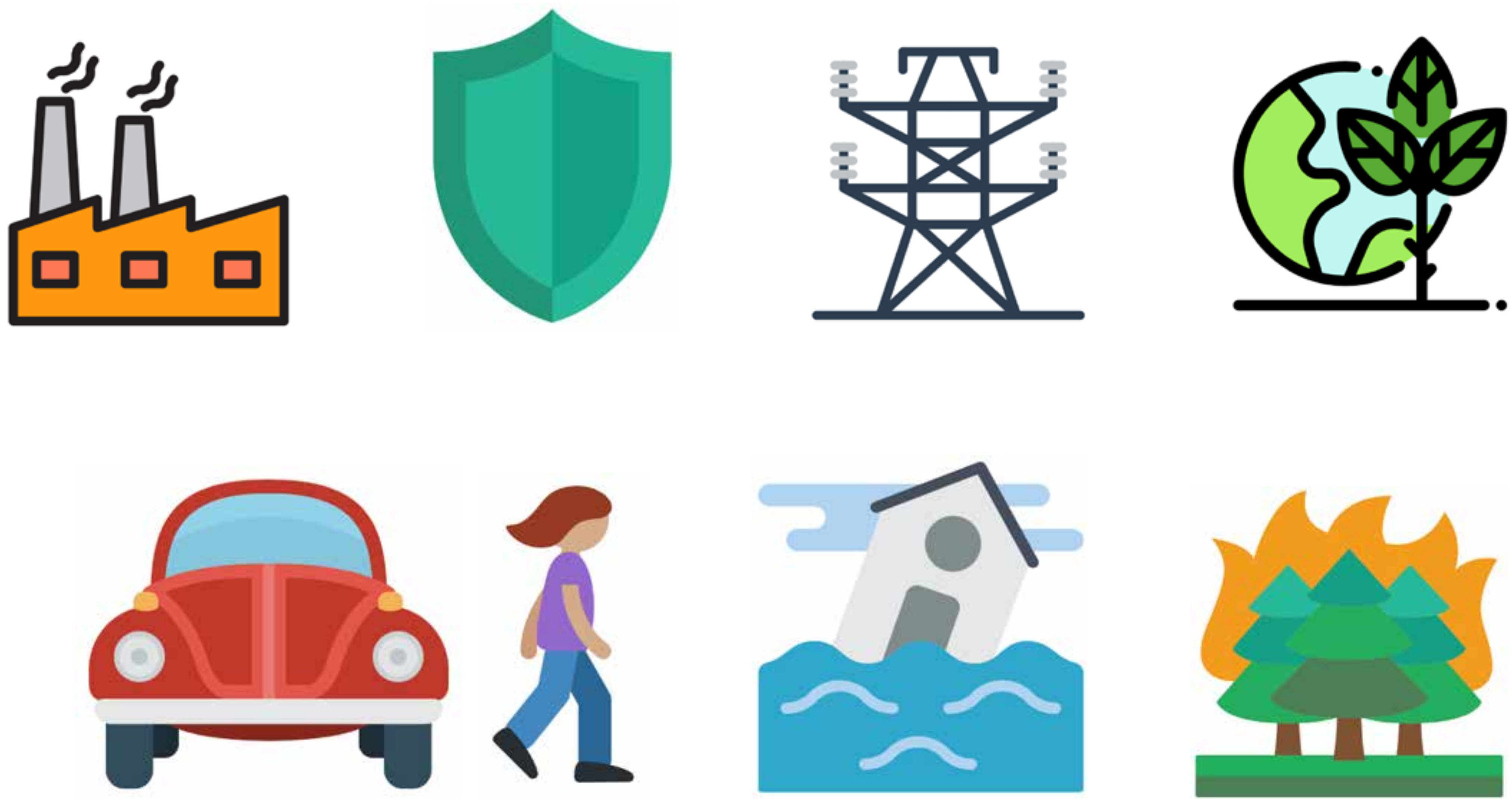
GRAPHCORE

MYTHIC

Intel Myriad

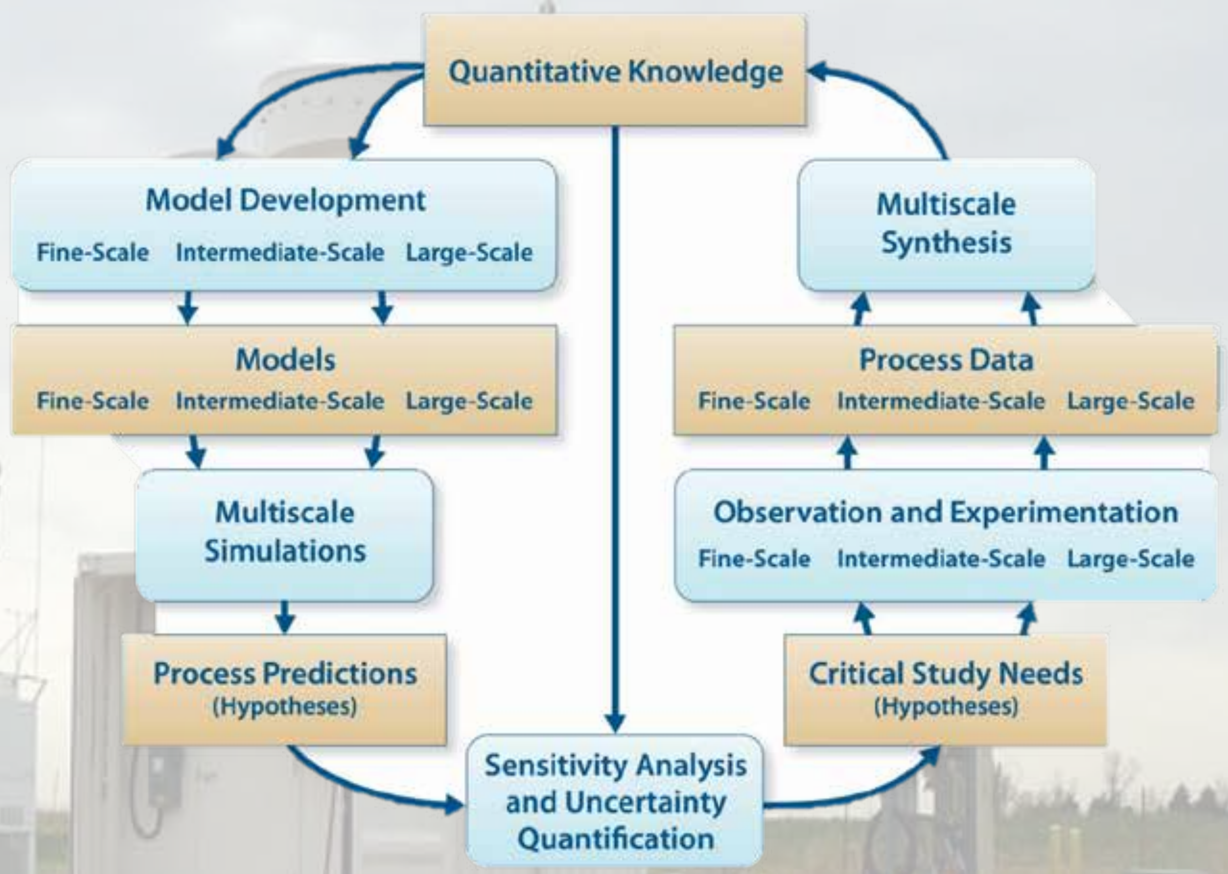
NVIDIA TX2

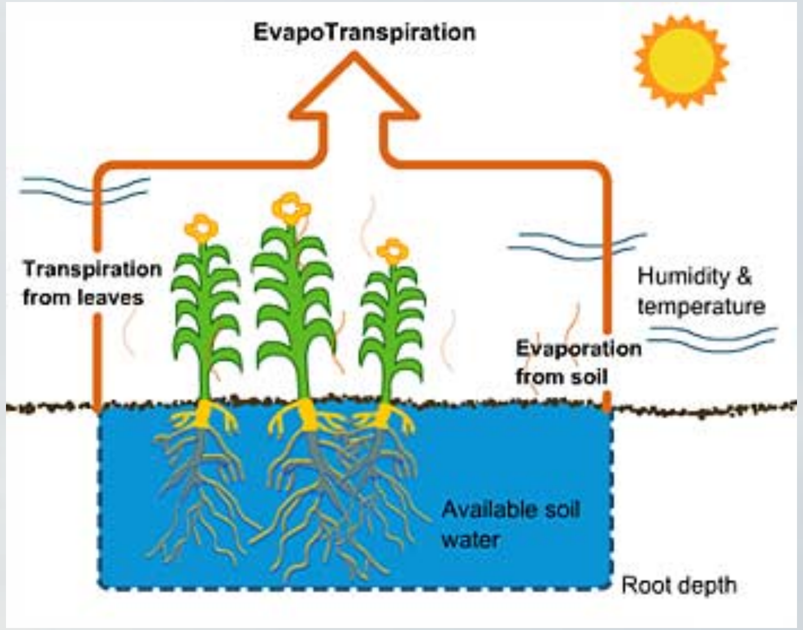
Are We Building a “Software Defined Instrument”?



SGP – a multiscale, integrated computational-experimental testbed

and a blueprint for other parts of the world





Harnessing The Computing Continuum

Science-driven Problems

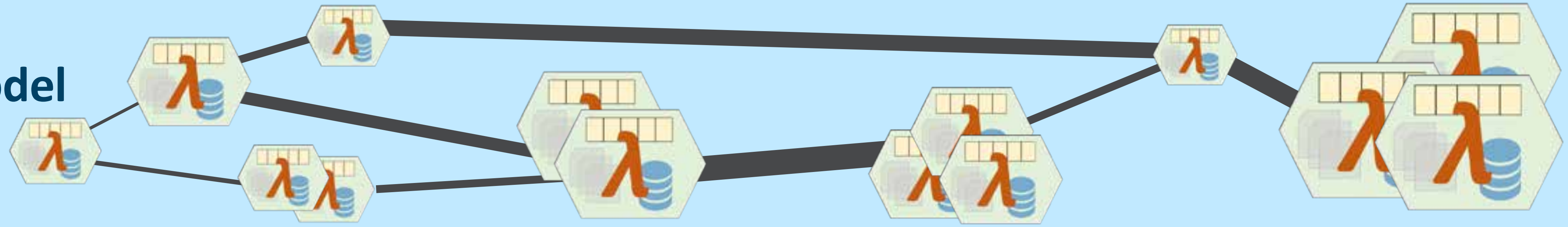
e.g.: “Predict urban response to rainfall, trigger intelligent reaction…”

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

Edgy Research: Edge-HPC

- Continually improving Edge-HPC Systems
  - Deep learning + lightweight training + continual improvement
  - Incremental model updates
  - Is Edge really a layer in the model?
- How will the OS/R and system software evolve for Edge-HPC?
  - Scheduling, security, resource management, streaming data
- Programming model & framework for Continuum Computing
- Optimized ML hardware for both Edge & HPC
- Theoretical foundations for failures and correctness of edge/training
- Dynamic resource management and adaptive inference priority
  - AI at the Edge is limited by power and computation – just like HPC
- **Fluid HPC** to support complex and on-demand workflows on future exascale

Questions?

Please drop by for my SC18 Talk

http://www.wa8.gl

http://arrayofthings.github.io

Thank You Funding:

- DOE EERE VTO
- Illinois DOT
- Exelon
- NSF
- CSIRO (in kind)
- DARPA (soon)
- ANL LDRD