AI @ Edge

IEEE CIC 2018 Tutorial

Mudhakar Srivatsa
Distinguished Research Staff Member
IBM TJ Watson Research Center

Data at the edge is causing us to rethink data

20 EB

per day of data generated at the edge

Data at the edge is causing us to rethink data

90%

Of data created over the last 10 years was never captured or analyzed

2x

Rate of data creation compared to the expansion of bandwidth over the past decade

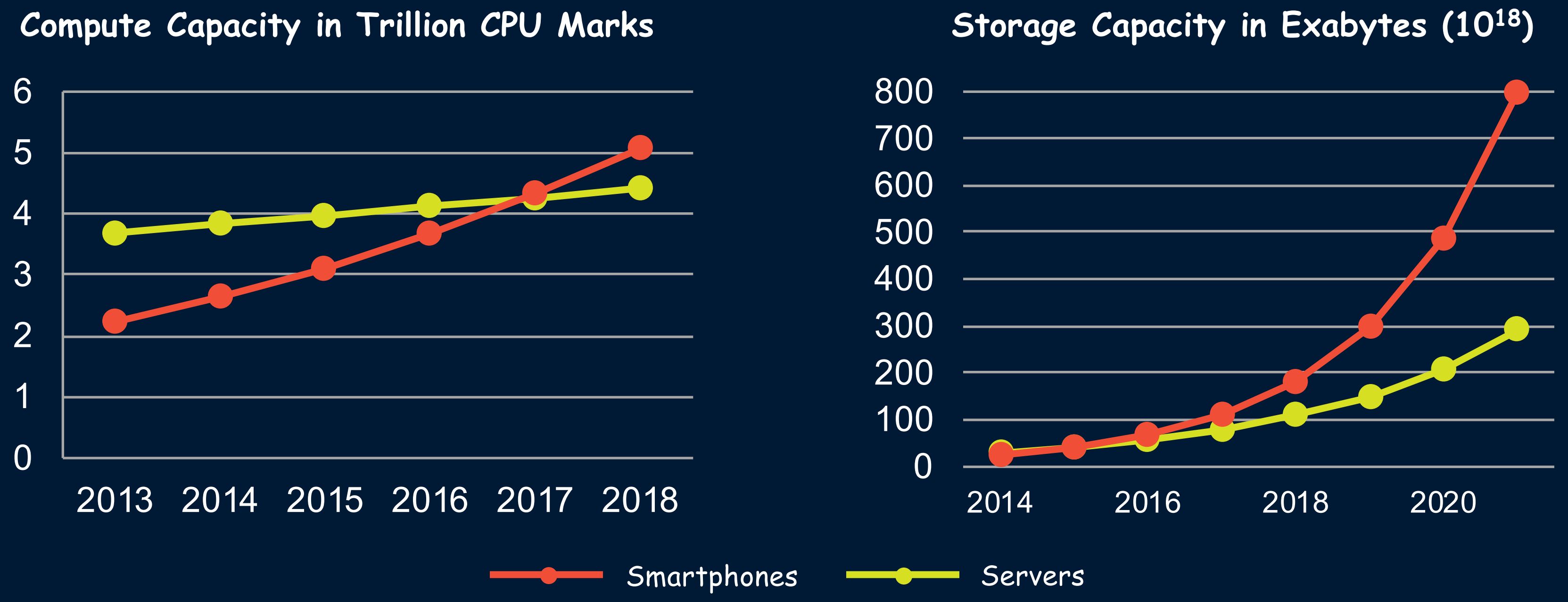
60%

Of valuable sensory data loses value in milliseconds

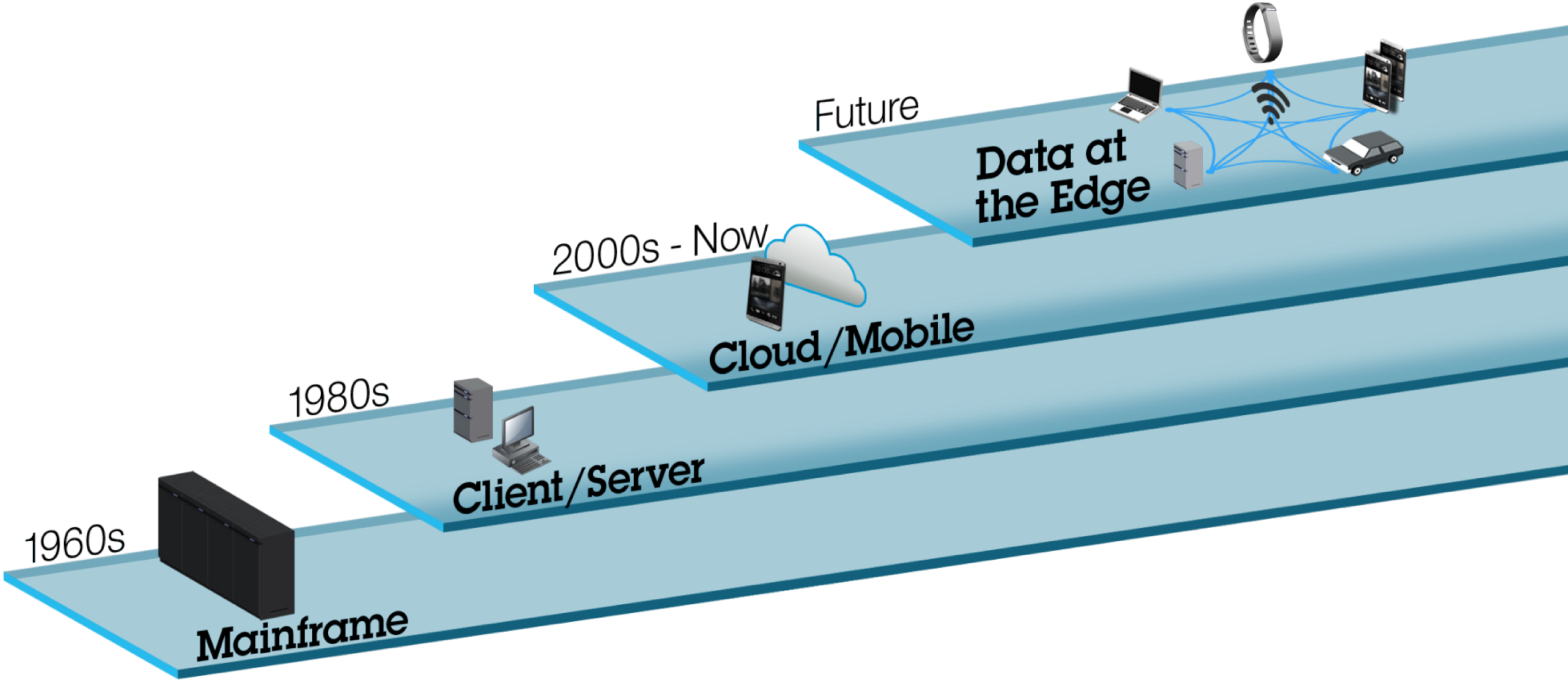
in 2017

The collective compute and storage capacity of smartphones surpassed all worldwide servers

Collective compute and storage at the edge exceeds that in the cloud



A new IT paradigm is emerging at the edge



IoT environment faces some key challenges

Bandwidth

Connectivity to cloud is too slow or intermittent

Regulations

Some data is restricted

Cost

Sending data to cloud is expensive

Privacy

Some data is too sensitive

Drivers for edge computing in IoT

IBM

Reduced Latency & Increased Local Control

- e.g. Vehicle-to-vehicle navigation and collision avoidance; make instant adjustments

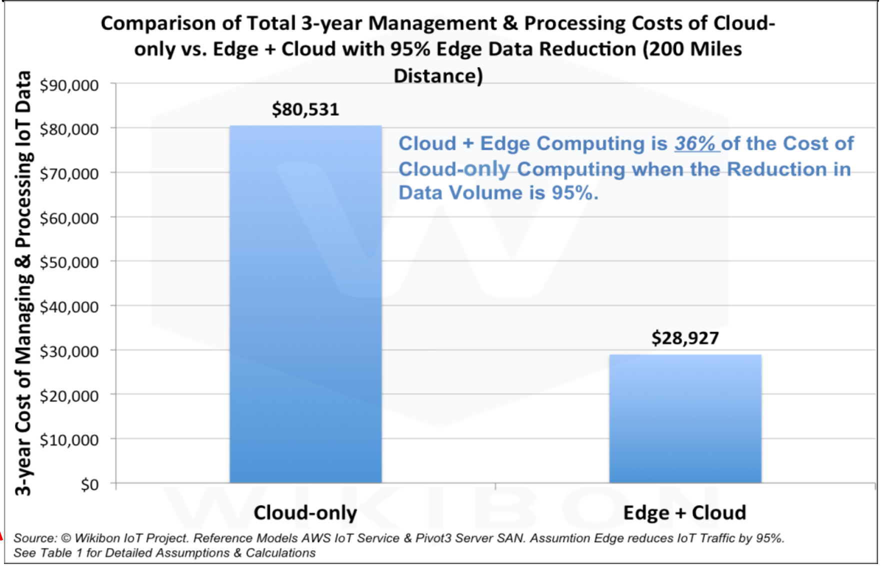
Optimization for Lower Costs

- e.g. large volume of data from oil rigs or video cams that's requires significant bandwidth and storage

Improved Security or Privacy

- e.g. Distributed risk in edge versus single point of failure in Cloud
- e.g. Localized scanning for early detection & mitigation of potential data breaches
- e.g. video surveillance data that cannot be saved

An independent research shows Edge + Cloud computing can significantly reduce costs over the Cloud-only option



Source: Frost & Sullivan, "Reaching out to the Edge: Defining distributed intelligence in IoT. Dec. 2015
http://wikibon.com/the-vital-role-of-edge-computing-in-the-internet-of-things/

Top Edge Computing Use Cases – Methodology for Selection

SOURCE DATA

46 IoT use cases*

Source: IDC Worldwide Semiannual Internet of Things Spending Guide

CRITERIA

1. **Require real time decision without latency**, e.g. Act on vehicles or planes in motion
2. **Large requiring significant bandwidth** or **too costly to transmit** back to cloud, e.g. Intelligent Oil field / drilling rig
3. **Security or privacy** requirements on the edge

Requires Edge
(11 use cases)

Some dependency on Edge
(11 use cases)

Edge not required
(24 use cases)

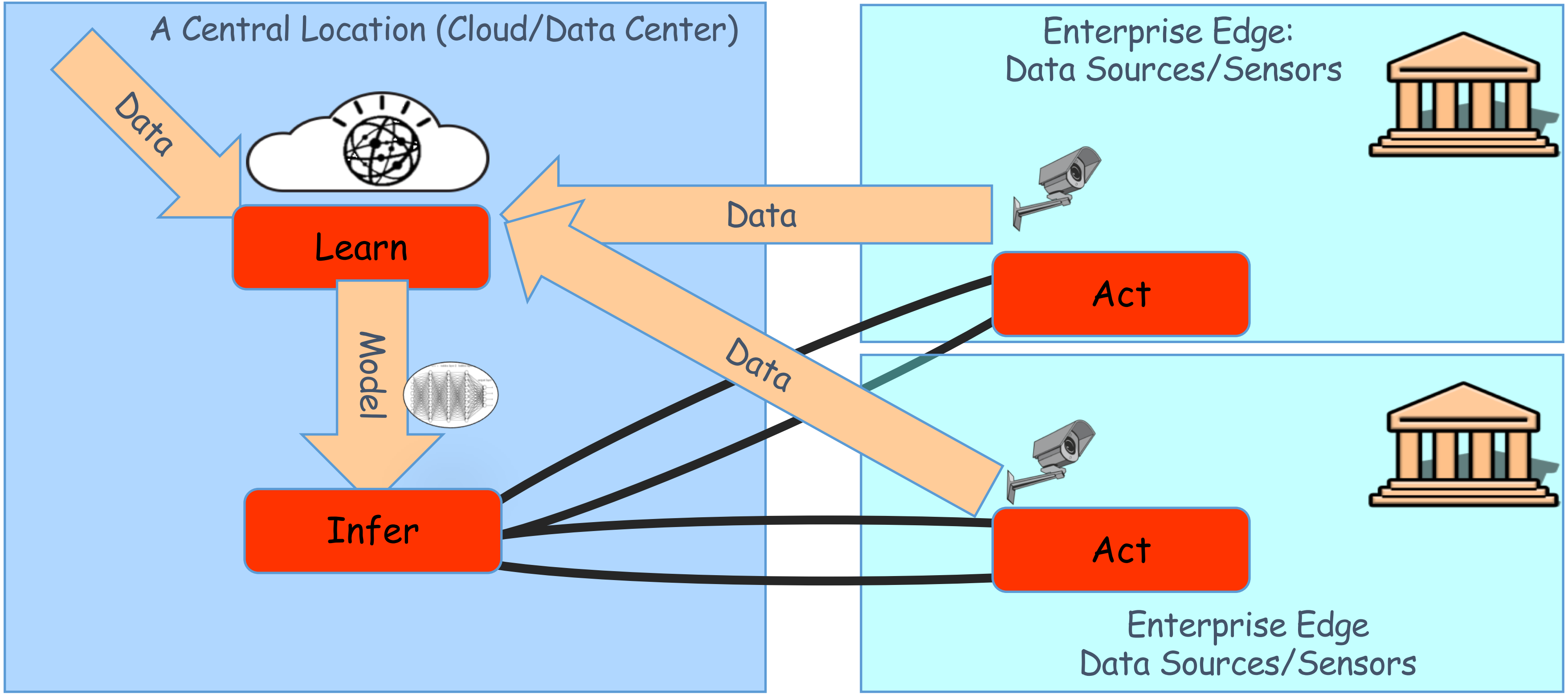
About half of IoT market-size requires edge or has a dependency on edge

AI @ Edge: Research Challenges

"Despite the power to process massive volumes of data and derive insightful insights, artificial intelligence applications have one major drawback - the brains are located thousands of miles away"

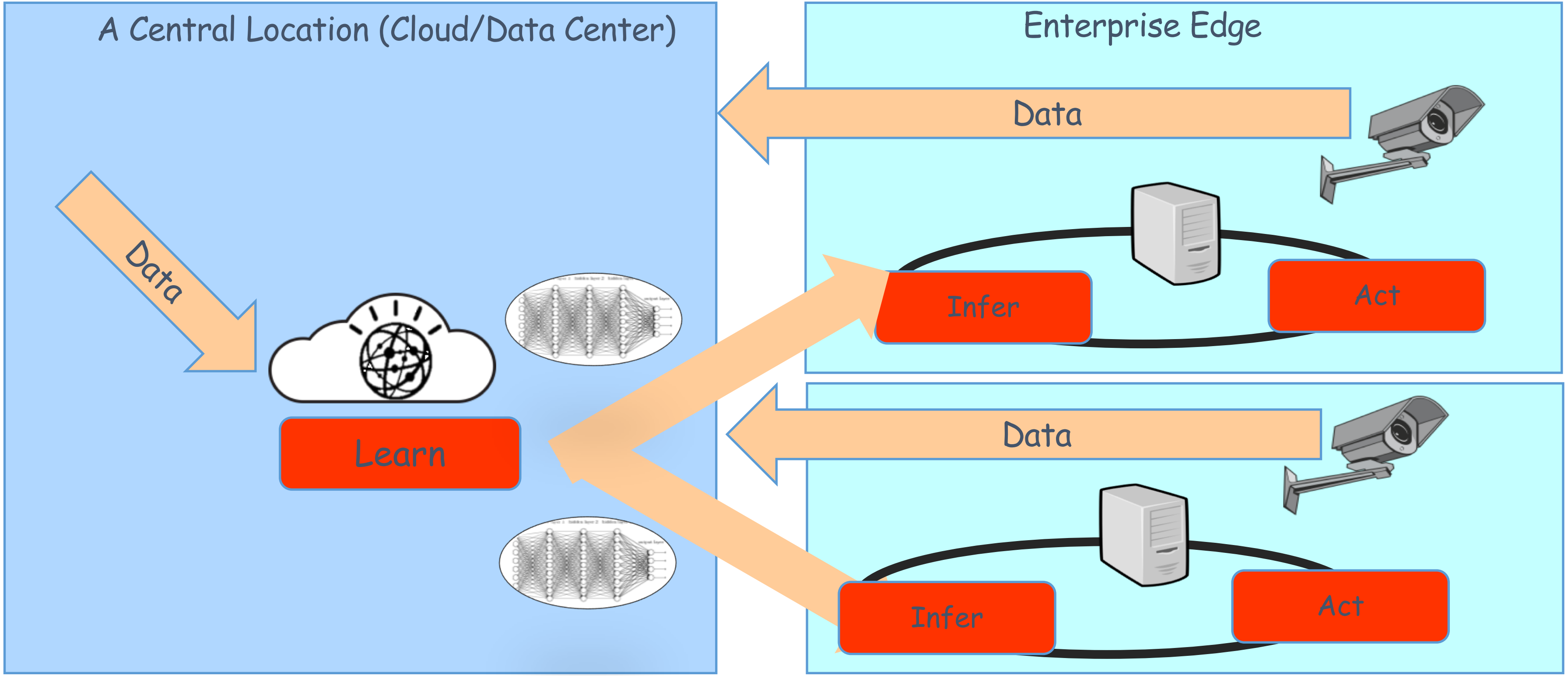
The current prevalent model for creating AI based solution

IBM



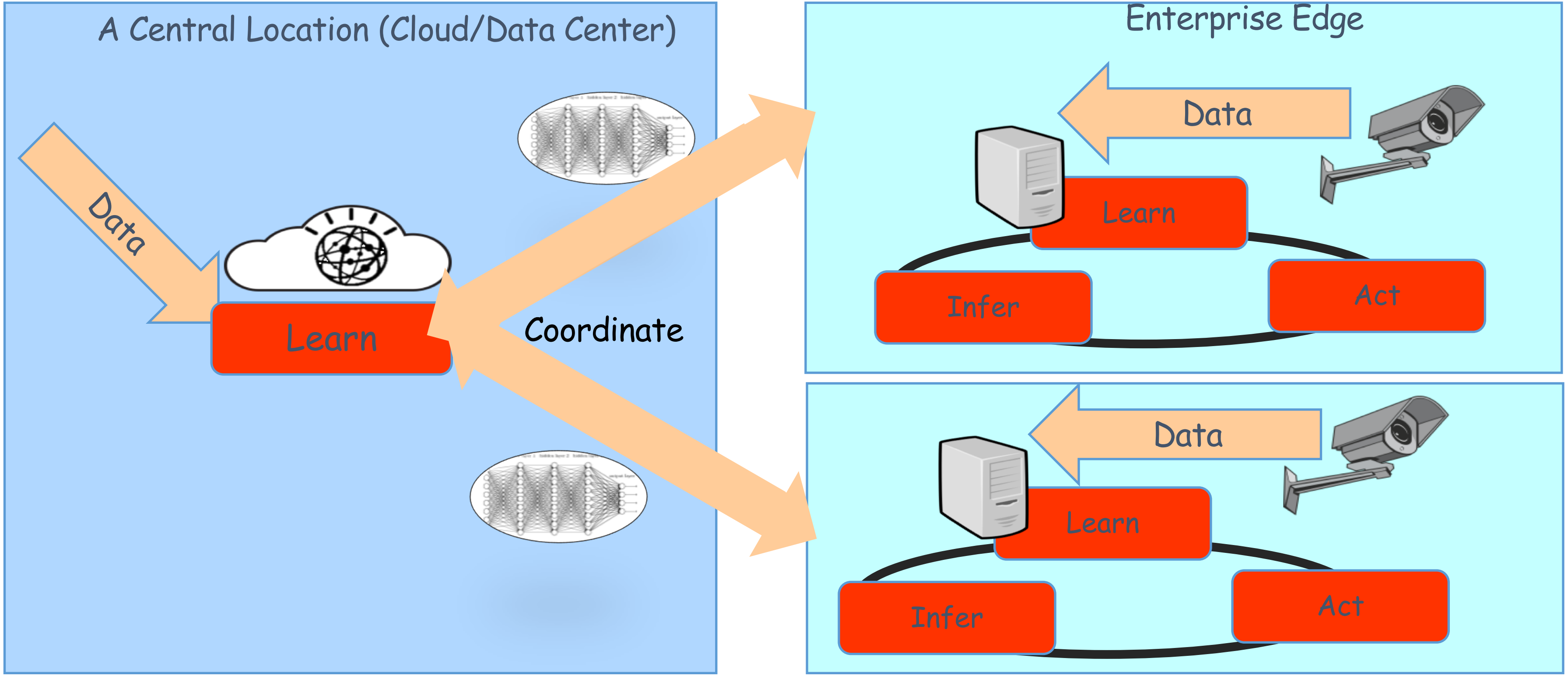
Regulations, Privacy Concerns, Network costs, Latency, Bandwidth Constraints are a hurdle for AI Solutions in many contexts.

AI @ Edge: Semi-distributed Model



As an intermediate stage, use the cloud to train the AI models, but move models out to the edge for inferences and action.

AI @ Edge: Fully Distributed Model

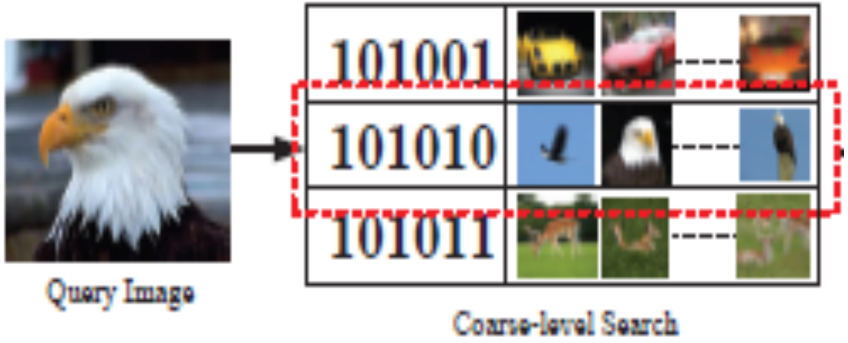


Learning happens at many different locations, and different locations coordinate the models they learn with each other

AI @ Edge Challenge I – Search Engine

- Deep learning models require large labeled training datasets: "small data" problem at the edge!
- Given a dataset, the first step is to bootstrap with a pre-trained model and customize this model for the given application: often manual, error prone and cumbersome

- There is no "search engine" for searching and ranking machine learning models for a given input dataset!
  - Ranking needs to capture partial match (match up to i^{th} layer), estimated cost of retraining (compute resources and labelled data requirement)
- Deep hash codes: a reduces the dimensionality of high-dimensional data by inducing hash collisions on similar inputs; use deep hash codes to fingerprint output (activations) from each layer in a trained network



Hash Code Method	Data Domain	Supervised
Latent Semantic Hashing [SH2009]	Text	No
Autoencoder [VLLBM2010]	Text, Images	No
Restricted Boltzmann Machine [TFW2008]	Text, Images	No
Tailored Feed-Forward Neural Network [MBBPS2014]	Text, Images	Yes
Deep Hashing [LLWMZ2015]	Image	No
Convolutional autoencoders [XPLLY2014]	Image	Yes
Deep Semantic Ranking Hash [ZHWT2015]	Image	Yes
Deep Neural Network Hashing [LPLY2015]	Image	Yes
Word2Vec [MCCD2013]	Text	No
Node2Vec [GL2016]	Graph	No

Ranking Machine Learning Models

IBM

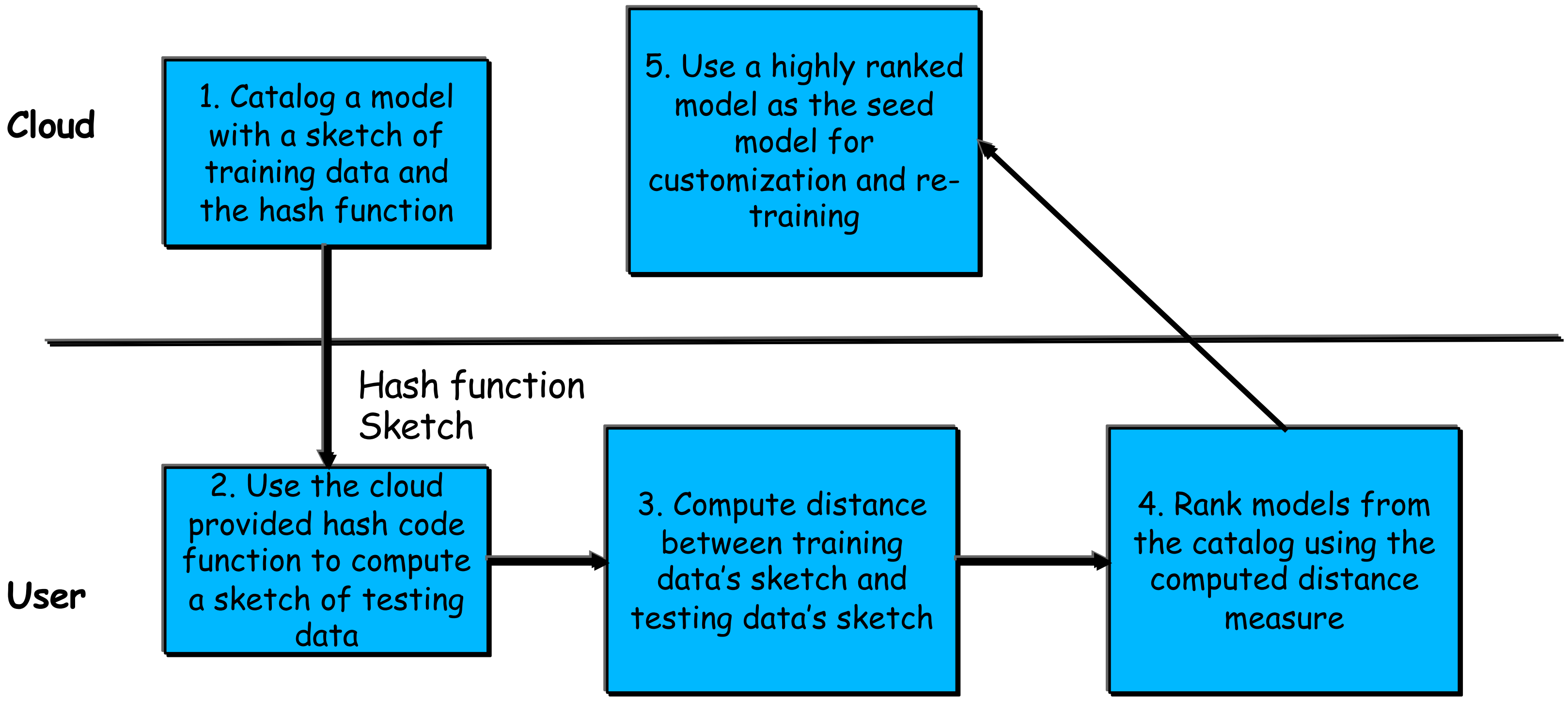
1. Compute a compact layer-by-layer setch of the trained model
   - For every training data x, compute clusters over h(x) (e.g., using k-means++ clustering)
   - Sketch: (c_i, w_i) where c_i is the i^{th} cluster head and w_i is its silhouette coefficient
   - Store the sketch and the hash function h along with the pre-trained model

2. Compute the sketch of the testing/input data
   - Same as (1) but seed the clustering algorithm with cluster heads obtained from (1)

3. For every pre-trained model in the catalog compute its rank using distance(w_i, w_i')
   - Sum of $(w_i - w_i')^2$ over all i (does not account for cluster size)
   - Wasserstein distance (Earth Mover Distance) to account for cluster sizes

4. Combine this score with a page-rank like score over the dependency graph of trained models
   - Edge (a → b): model b was retrained from model a, and the weight of this directed edge is obtained from step 3

Process Flow

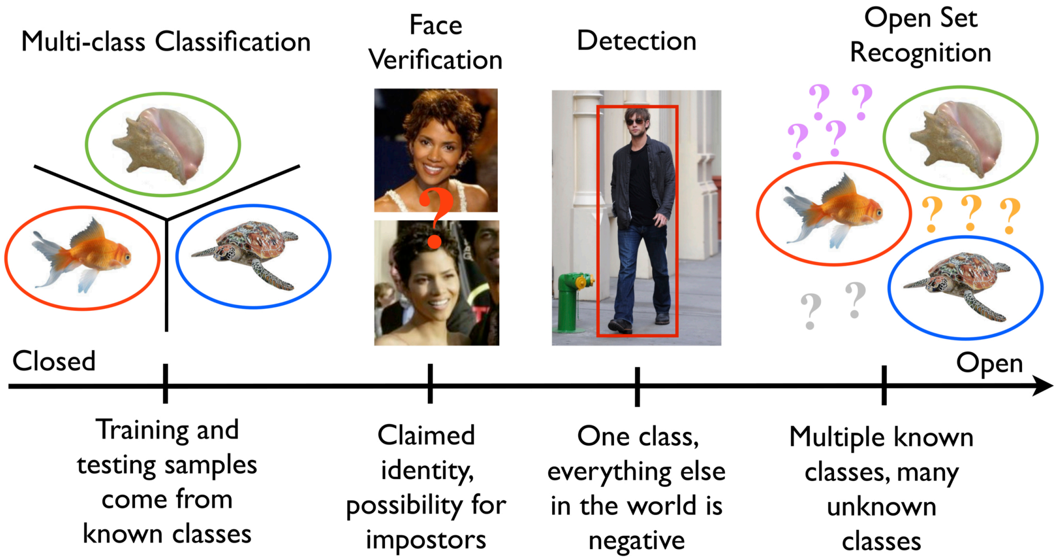


Multiple realizations of the process flow are possible: above shows a workflow where the training data is never released to the user (only its sketch is shared) and the testing data is held private until a suitable model is discovered in the catalog

AI @ Edge Challenge I – Open Sets Problem

IBM

- Typical setting: model is trained at the cloud using labeled dataset; the trained model is scored at an (unattended) edge
- Adapt and customize a pre-trained model at an edge
  - Anomaly detection: check is an unseen unlabeled input at the edge is anomalous
  - Open set problem: detect a novel class at this edge



Courtesy: https://www.wjscheirer.com/projects/openset-recognition/

Anomaly and Open Sets Detection

- During model training phase on labeled data (at cloud), compute a model sketch
  - Class-wise centroids with normalized inertia measures at each layer of the network
  - Normalized inertia:
    - $I(C,X) = \sum_{i=1}^{N} \frac{||C^{*}-X_i||_2^2}{N}$,
    - where $C^{*} = argmin_C \, ||C - X_i||_2$

- During model scoring phase on unlabeled data (at edge) compute distance between data at every layer and the model sketch

- Anomaly scores and open set characterization using: silhouette coefficients and Wasserstein metric (Earth mover distance)

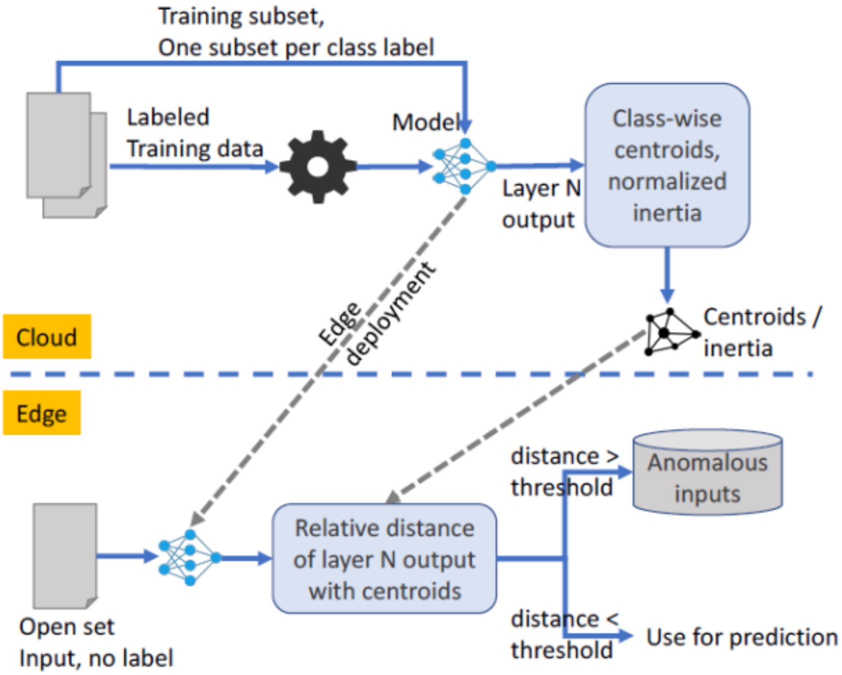
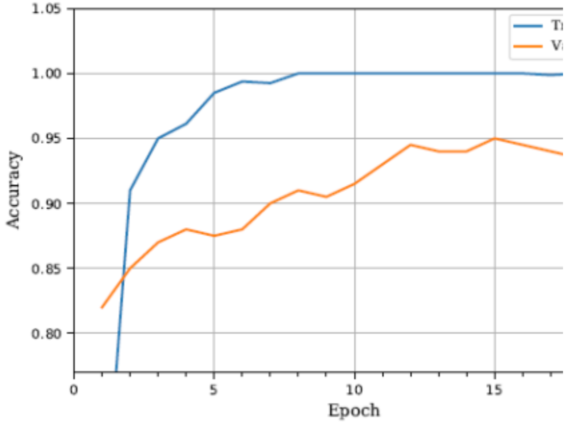
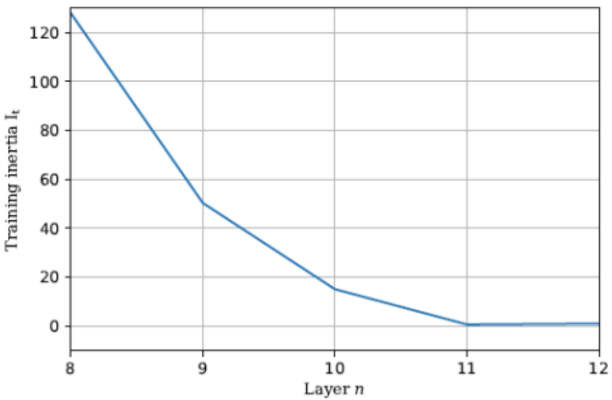


Illustration on MNIST dataset

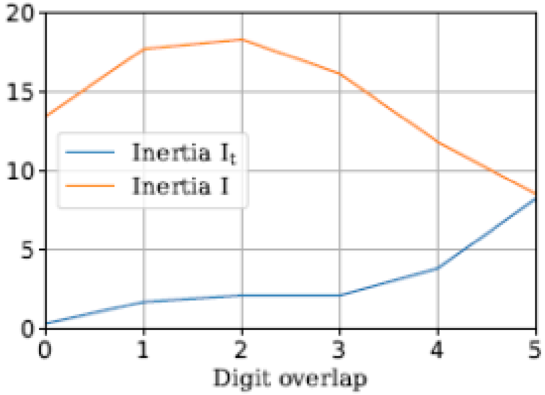
Data is partitioned into two classes: (0-4) for training and (0-9) testing classes



Training and validation accuracy over epochs



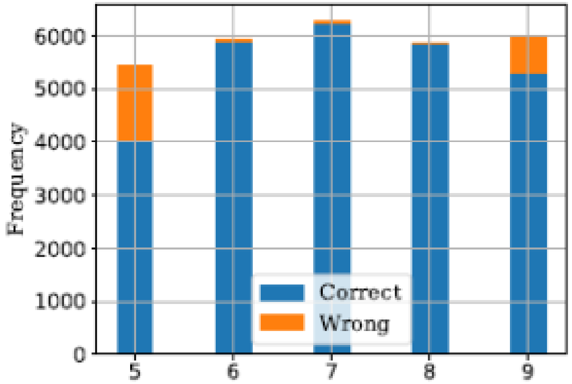
Normalized inertia at different layers for training data



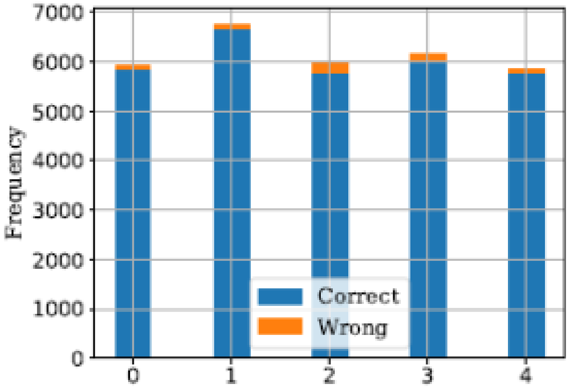
Normalized inertia at layer 12 with novel classes (I_t: training; I: testing)

Anomaly Detection and Open Set Recognition

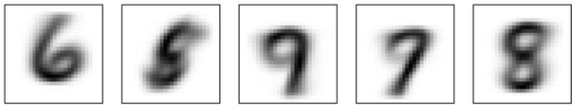
IBM



Anomaly detection accuracy on open set inputs



Anomaly detection accuracy on training data (closed set inputs)



Centroids over anomalous inputs (shows high confusion for 5 and 9)

https://dataplatform.cloud.ibm.com/analytics/notebooks/v2/1c9fa74e-55bb-4407-9045-4f0f5a12b47d/view?projectid=6b12966f-3621-4f15-bbfe-1f79fb5659fe&context=analytics

Application I: Customizing speech-to-text models at edge

IBM

- Typical speech-to-text models are trained on news corpus and lack customization to specific industry domain
  - On noisy input the output had lot of low confidence transcriptions to Saddam Hussein, Iraq, etc.

- Customization process involves identifying errors in the output and correcting the models. Error correction typically happens through manual feedback

- Apply deep hash codes to the output of speech-to-text and obtain anomalous clusters (e.g., why, fly – which are both incorrect transcriptions of WiFi)
  - The novel word “WiFi” can now be added to the speech-to-text model → example of model customization with limited supervision at the cloud/edge

- Result: output accuracy improves from a baseline of 71% to 89%

Application II: Fingerprinting IoT Devices

- Examine DNS (Domain Name Service) requests from a device to classify it as IoT vs. non-IoT; if IoT identify a more specific device type (e.g., camera, LIFX bulb, Wemo switch, etc.)
  - Reduce error rate to 0.21% from 4.22% (20x improvement)

DNS name	Three Most Similar DNS names (DNS name, cosine similarity)		
chat.hpeprint.com	h20593.www2.hp.com, 0.75	xmpp006.hpeprint.com, 0.72	h10141.www1.hp.com, 0.68
0.invoxia.pool.ntp.org	sip.invoxia.com, 0.97	icecast.icecast.sbs.com.au, 0.93	ws.invoxia.io, 0.93
r1—sn-p5qlsnez.googlevideo.com	r9—sn-p5qlsney.googlevideo.com, 1.0	vassg142.ocsp.omniroot.com, 1.0	gv.symcd.com, 0.96
pscfcb6ec6.pubnub.com	pscab6d5d1.pubnub.com, 1.0	psc3f5c69e.pubnub.com, 1.0	psc8a67f35.pubnub.com, 1.0
v4.netatmo.net	_vpn._udp.netatmo.net, 0.98	v3.netatmo.net, 0.97	v5.netatmo.net 0.97
ntp1.glb.nist.gov	time.nist.gov.lan, 0.70	time.nist.gov, 0.46	time1.google.com, 0.33

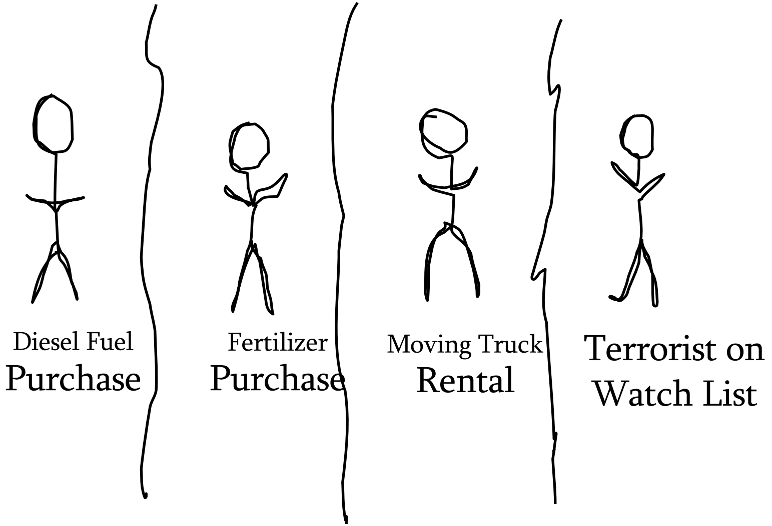
Hash codes similarity on DNS requests from non-IoT devices

DNS name	Three Most Similar DNS names (DNS name, cosine similarity)		
newyorker.com	buzzfeed.com, 0.96	nytimes.com, 0.95	nymag.com, 0.95
nba.com	vividseats.com, 0.98	theundefeated.com, 0.98	espnfc.us 0.98
sharelatex.com	overleaf.com, 0.96	slack-imgs.com 0.96	slack-edge.com, 0.96
sinovision.net	asiancc.net, 0.96	hking.hk, 0.96	uschinapress.com, 0.96
247checkers.com	cardgamesolitaire.com, 0.99	123freecell.com, 0.99	solitairetime.com, 0.99
akamaiedge.net	akadns.net, 0.99	collabserv.com, 0.99	akamai.net, 0.99

Hash codes similarity on DNS requests from IoT devices

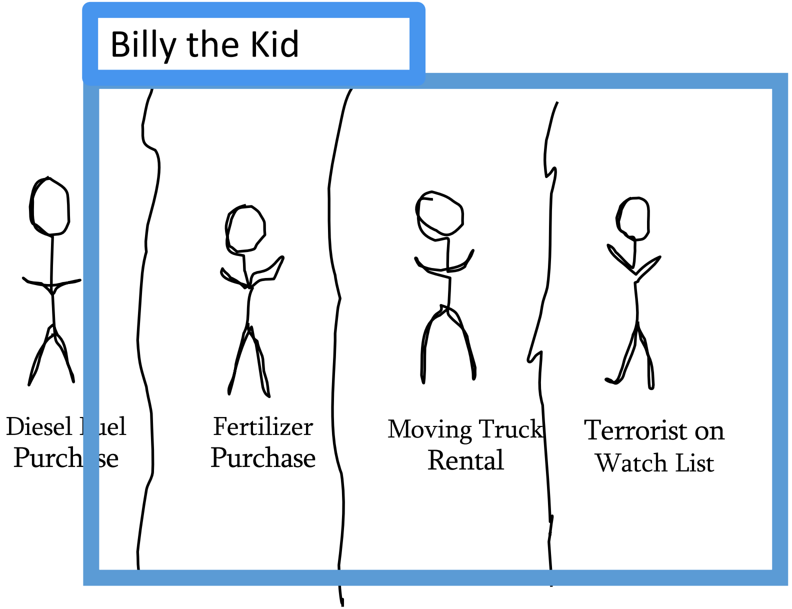
Case Study I: Maritime Piracy and Drug Trafficking

AI @ Edge: Maritime Piracy and Drug Trafficking



Channel Separation

AI @ Edge: Maritime Piracy and Drug Trafficking



Channel Consolidation

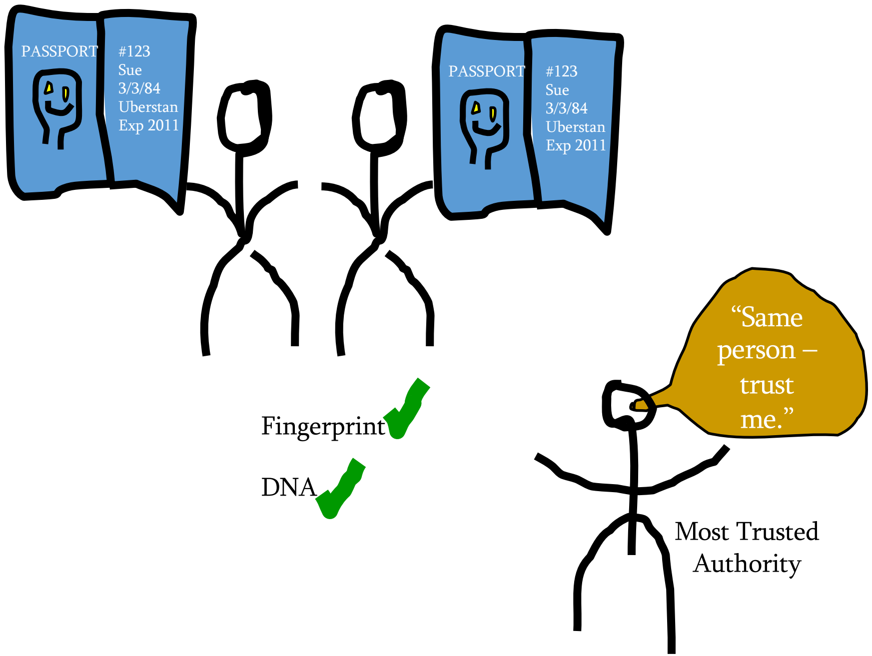
Entity Resolution is Essential for Prediction

- Is it 5 people each with 1 account or is it 1 person with 5 accounts?
- Is it 20 cases of Ebola in 20 cities or one case reported 20 times?

Re-thinking Entity Resolution

People	Cars	Router
Name	License Plate No.	Serial Number
Address	VIN	MAC Address
Date of Birth	Make	IP Address
Phone	Model	Make
Passport	Year	Model
Nationality	Color	Firmware Vers
Biometric	Etc.	Etc.
Etc.		

Consider Lying Identical Twins



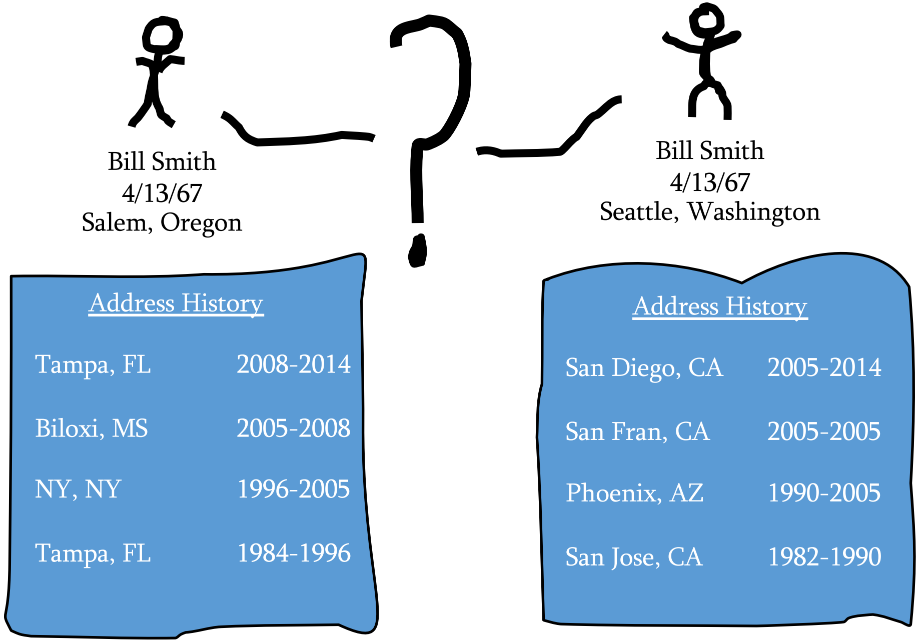
D'oh!

The same thing cannot be in two places at the same time

D'oh!

People	Cars	Router
When Name	When License Plate No.	When Serial Number
Where Address	Where VIN	Where MAC Address
Date of Birth	Make	IP Address
Phone	Model	Make
Passport	Year	Model
Nationality	Color	Firmware Vers
Biometric	Etc.	Etc.
Etc.		

Life Arcs are Telling



Multi-Resolution Life Arcs for Anomaly Detection

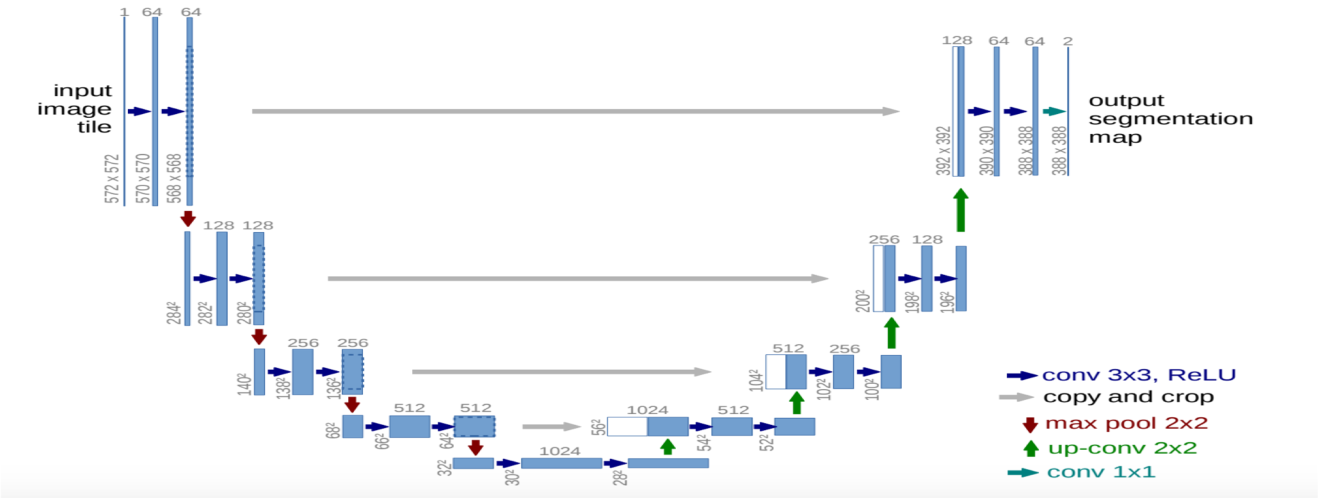
$$hash(40.00105, -78.30105) = dr07d1yzj21$$
$$hash(40.001, -78.301) = dr07d1yy$$
$$hash(40.01, -78.2) = dr07se$$
$$hash(40, -78) = dr0e$$

Index name	Deterministic	Extensible	Uniform	Bitwise
Grid	✓	X	Unbounded	X
Quad-tree	X	✓	4x	X
KD-tree	X	✓	dx*	X
R-tree	X	✓	1x	X
Geohash	✓	✓	1-2x	✓

- Efficiency gains with increasing cost ($$$)
  - 2x in software
  - 20-50x with FPGA/GPUs
  - 1000x with TCAMs

When Life Arcs are Missing…

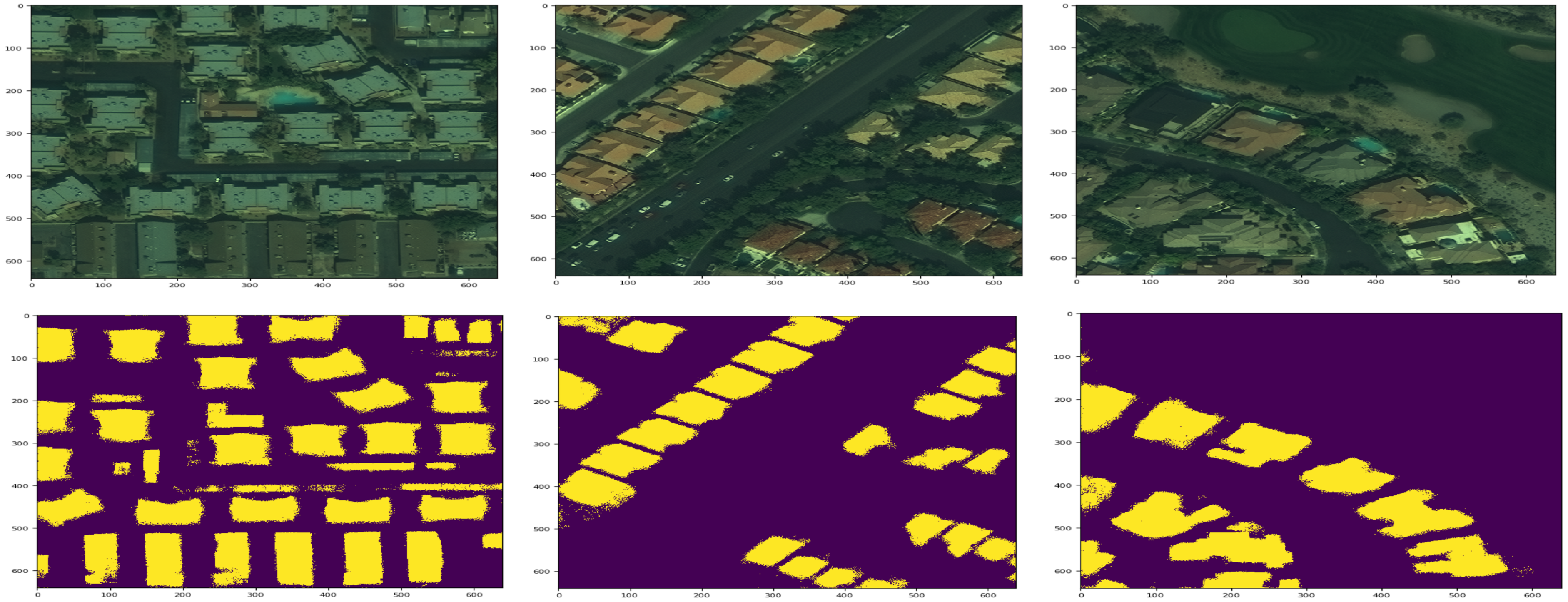
- Deep Learning models over low-orbit satellite imagery
  - Convolutional autoencoder-decoder pipeline to obtain a binary segmented 1-channel image from a 3-channel input image
- A modified U-Net pipeline (proposed initially for biomedical image segmentation)
  - Modifications: loss function optimized for improving IOU (Intersection Over Union) metrics, number of levels, convolution kernel sizes



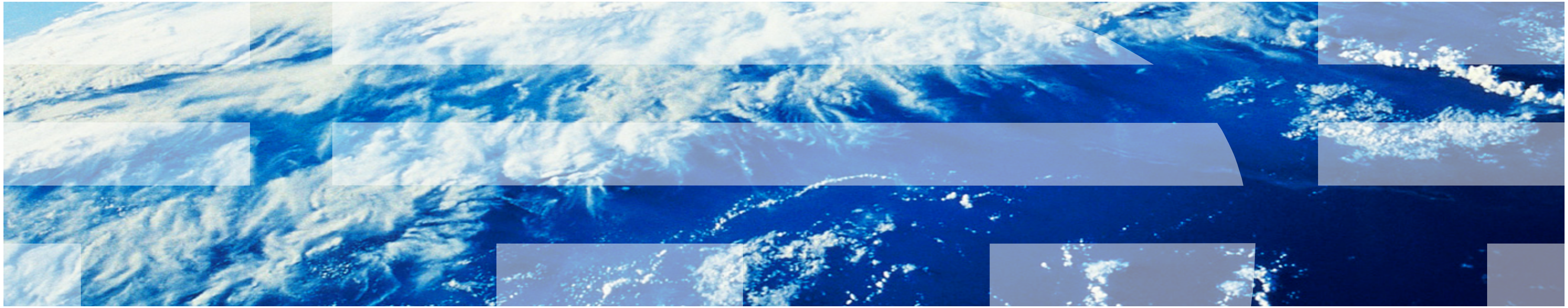
U-Net Architecture

Building Rooftop Extraction Results

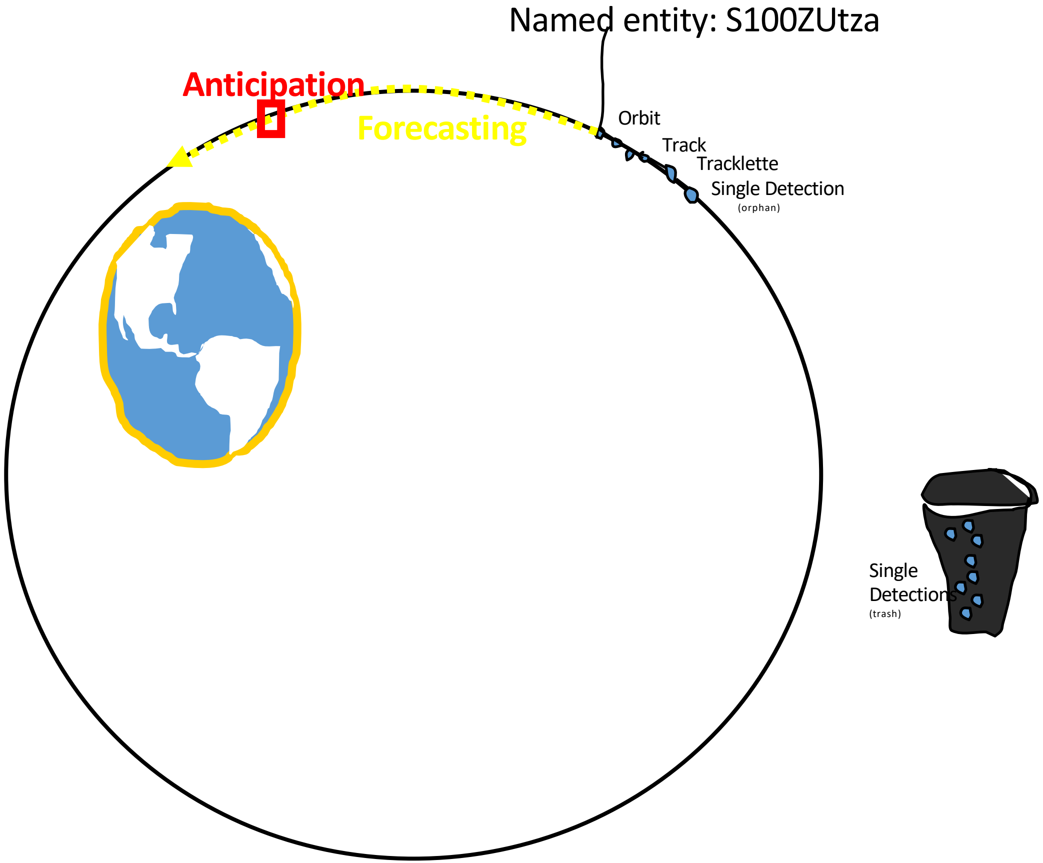
- Training Data: SpaceNet Buildings Dataset, containing data from Paris, Shanghai, Las Vegas, Khartoum and Rio de Janeiro (~10K images)
- **IOU: 0.81; Accuracy: 0.98**



Asteroid Hunting



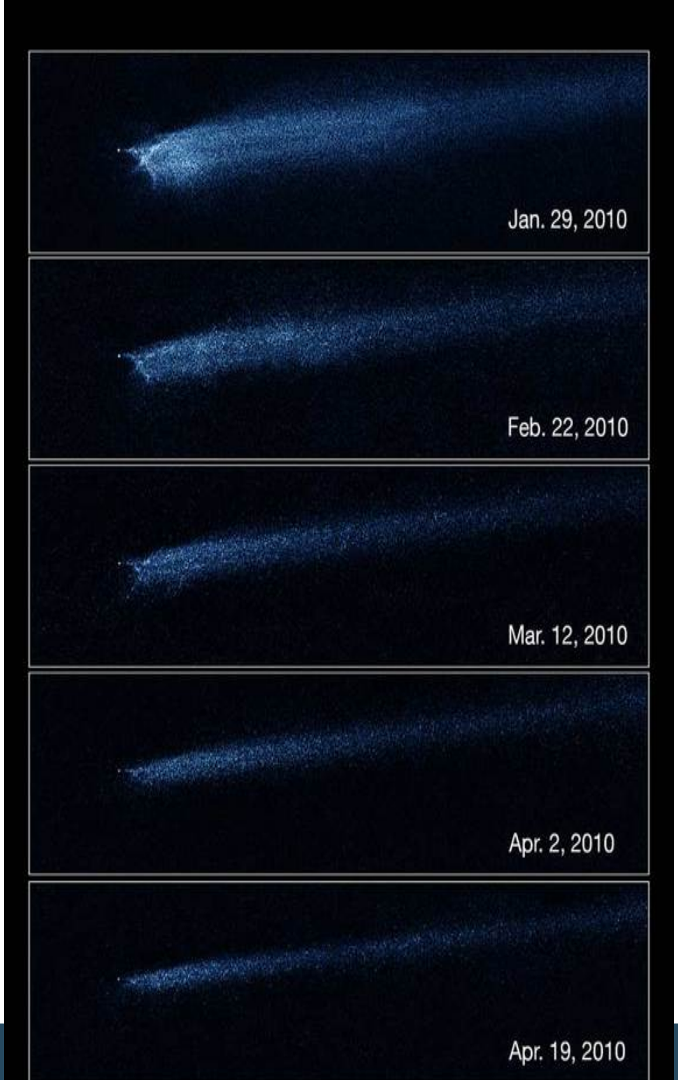
From Orphans to Orbits



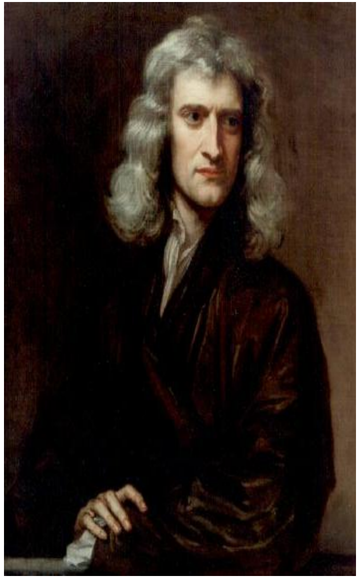
Asteroid-Asteroid Encounters

"We have directly observed a [collision between asteroids] for the first time, instead of having to infer that they happened from million-year-old remains."

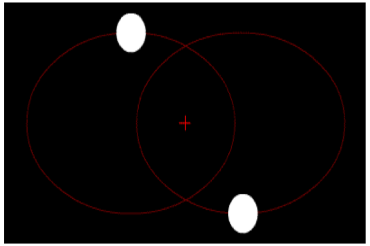
Colin Snodgrass
Planetary Scientist
Max Planck Institute for Solar System Research



Two-body Problems are easy to solve

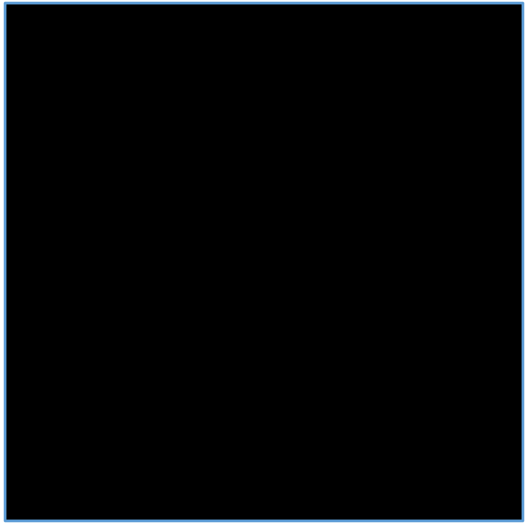


Isaac Newton

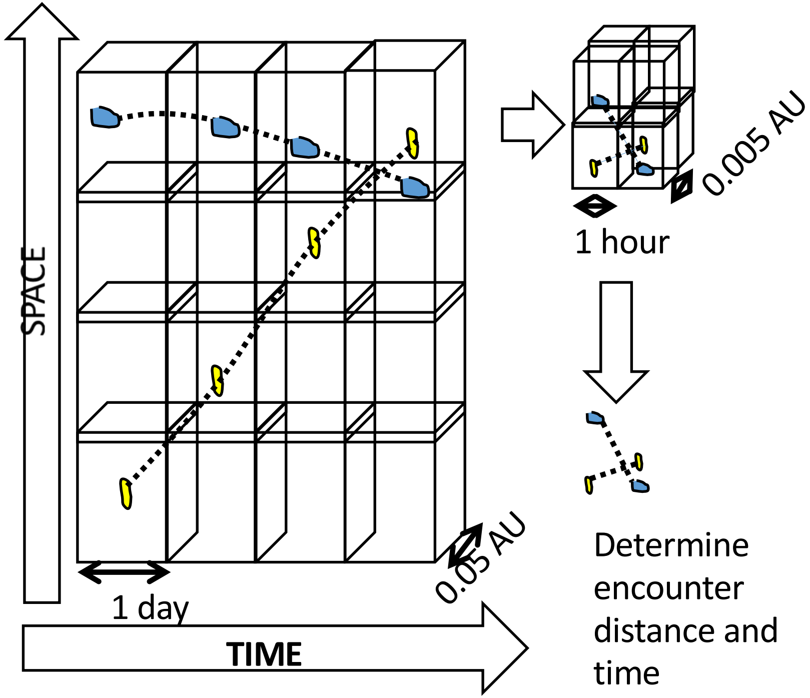


N-body Problems are hard!

Pierre-Simon Laplace



3D Life Arcs


SPACE
TIME
1 day
0.05 AU
1 hour
0.005 AU
Determine encounter distance and time

600K Asteroids x 25 years

Encounters by Proximity

Encounter	Distance	Asteroid 1	Size	Asteroid 2	Size
May 1, 2032 63353.9318 (MJD)	299km 0.000002 (AU)	00A9170	2-4km 15.8 (H)	0008758	4-9km 13.9 (H)
Nov 24, 2016 57716.07911 (MJD)	449km 0.000003 (AU)	00P5634	1-2km 17.4 (H)	0055711	2-5km 15.5 (H)
Jan 11, 2018 58129.29692 (MJD)	449km 0.000003 (AU)	K08E88J	530-1200m 18.3 (H)	00N0062	2-4km 15.8 (H)

Encounters by Size

Encounter	Distance	Asteroid 1	Size	Asteroid 2	Size
Feb 18, 2028 61819.1561 (MJD)	70K km 0.000469 (AU)	0000346	110-240km 7.13 (H)	00A4356	2-5km 15.5 (H)
Feb 28, 2031 62925.12725 (MJD)	54K km 0.000359 (AU)	0000348	35-75km 9.4 (H)	00G7226	2-4km 16.1 (H)
Oct 25, 2036 64991.01073 (MJD)	43K km 0.000289 (AU)	0000690	65-150km 8.02 (H)	0083174	3-7km 14.3 (H)

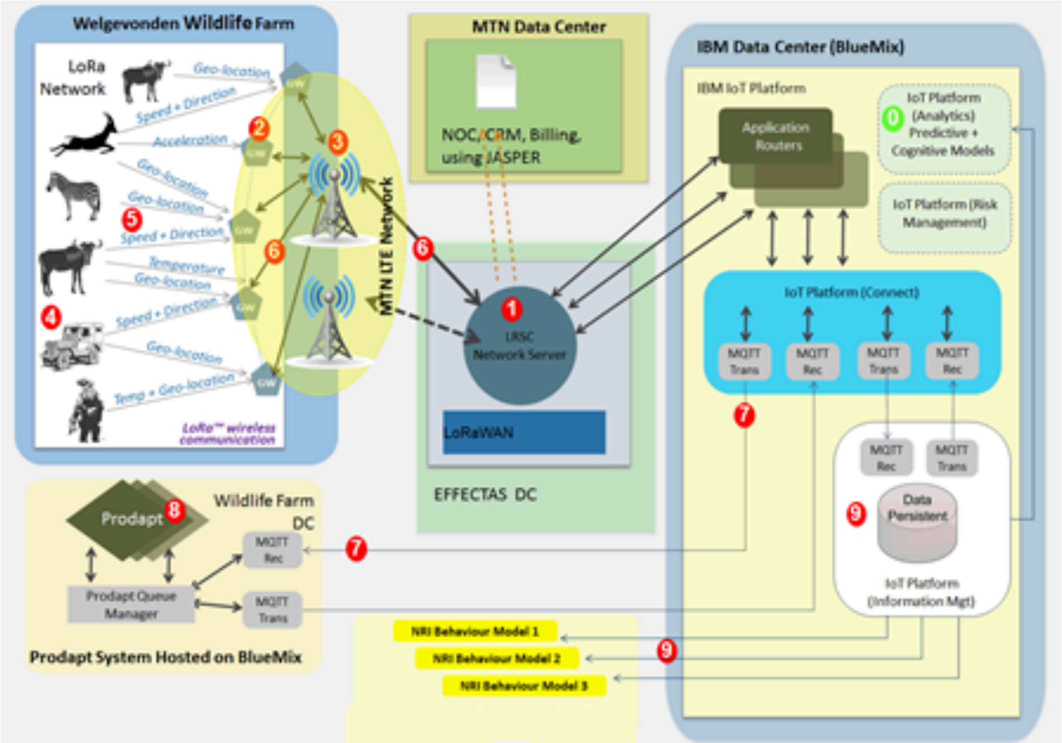
Orders of magnitude improvement in performance

Supports incremental addition of newly discovered asteroids

A few predictions validated by Univ of Hawaii telescope

Case Study II: Protecting Rhinos at Welgevonden Game Reserve, South Africa

AI @ Edge: Protecting Rhinos at Welgevonden Game Reserve, South Africa



Tag is applied to non-endangered species (applying them on Rhinos will allow them to be triangulated by poachers)

Learn predator vs. poacher pattern from sensor data:

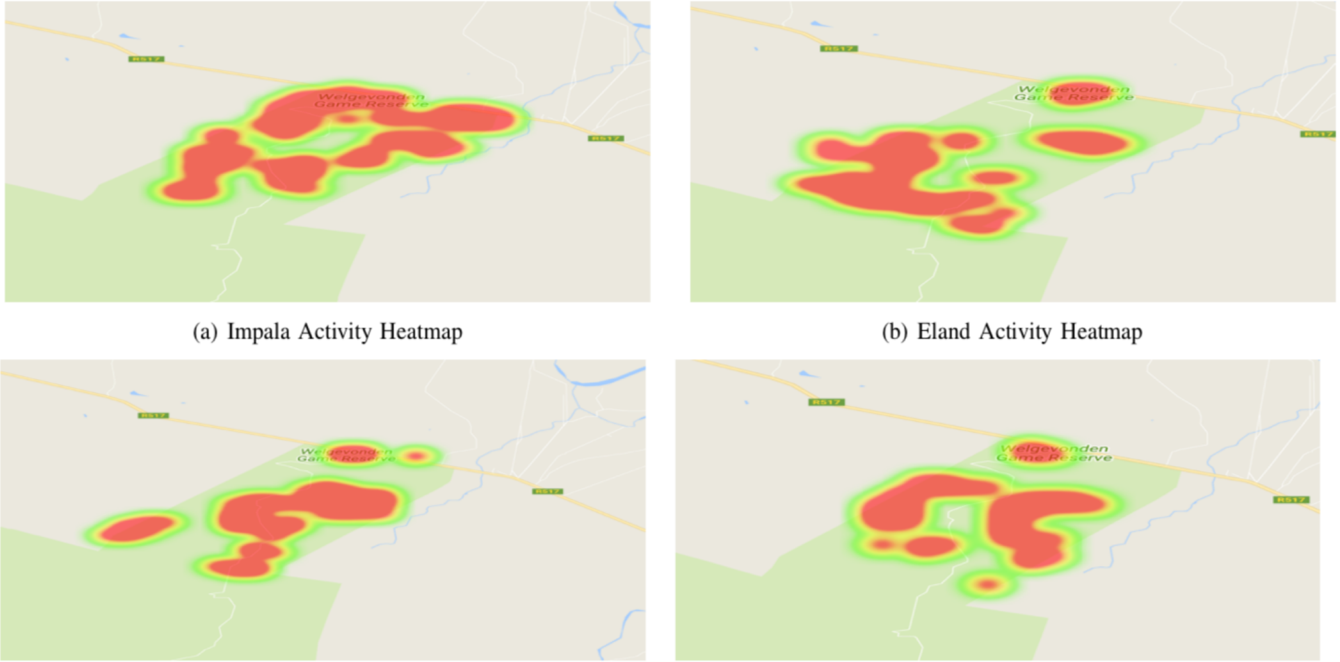
- Per-animal models identify anomalies (but cannot distinguish between predator and poacher)
- Group models (scatter patterns) distinguish between predators and poachers

IBM press release: https://www.ibm.com/thought-leadership/smart/
Bloomberg: https://www.bloomberg.com/news/articles/2017-09-19/mtn-ibm-to-combat-rhino-poaching-with-collars-for-prey-animals
Economist: https://www.economist.com/special-report/2017/11/09/electronic-surveillance-may-save-the-rhino
Youtube video: https://www.youtube.com/watch?v=E9olFUDD_2M

Coarse Grained Patterns

- Data collected from animal collars stored in DashDB
- Data from 112 collars fitted on: Impalas, Zebras, Wildebeests, Elands
- Data types:
  - **Latitude/Longitude (GPS)**
  - **Accelerometer**
  - Magnetometer
  - Temperature

Approach: Spatiotemporal clustering



(a) Impala Activity Heatmap

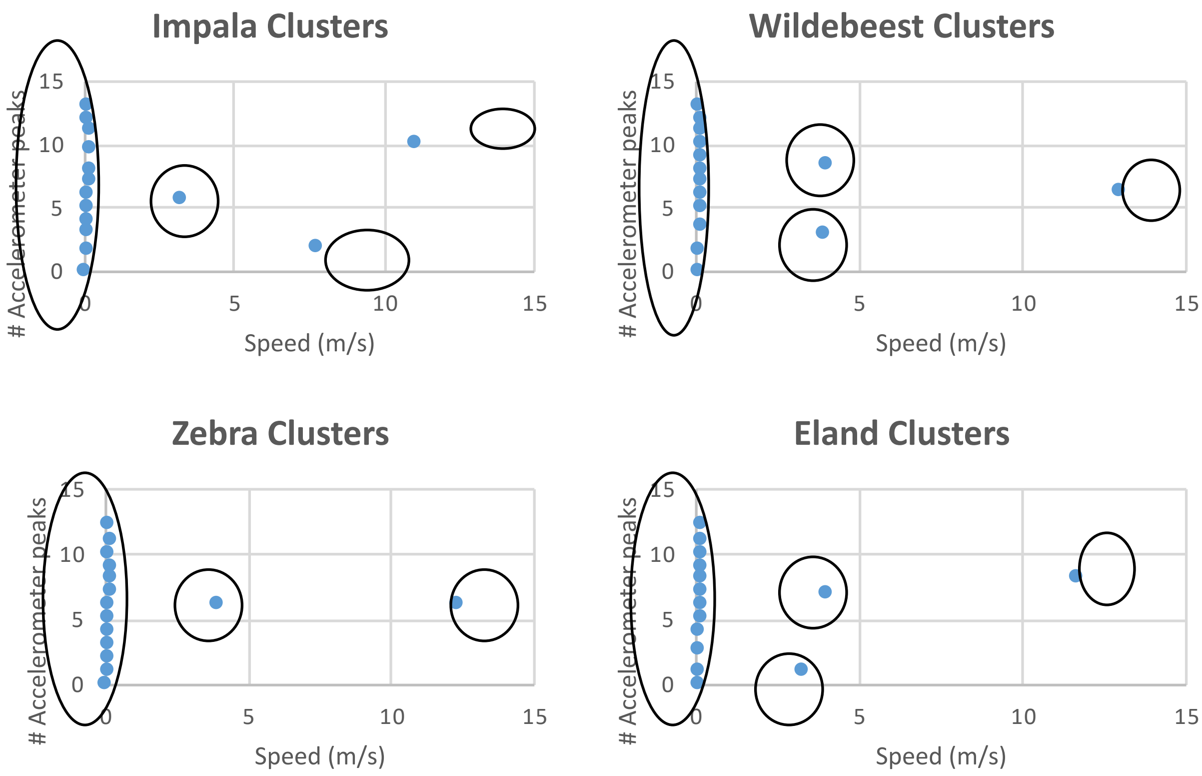
(b) Eland Activity Heatmap

(c) Wildebeest Activity Heatmap

(d) Zebra Activity Heatmap

Heatmap Activity for Different Animal Species During Morning Hours (Single D

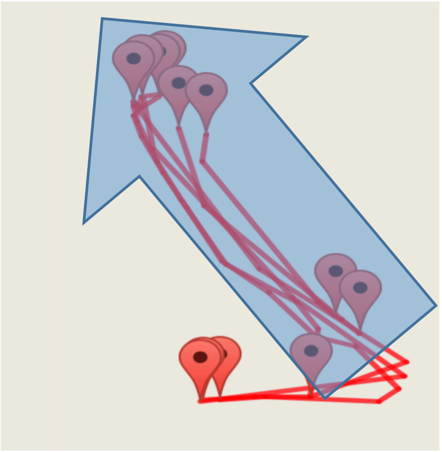
Unsupervised Pattern Learning

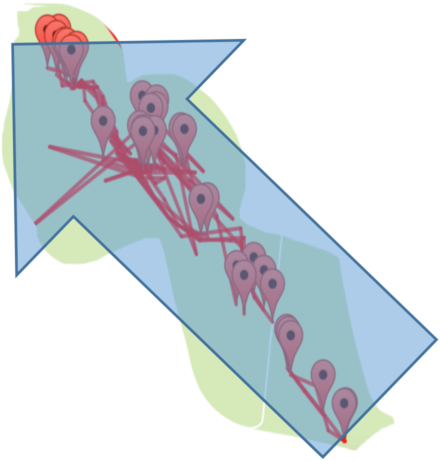


K-means Clustering, k = 15

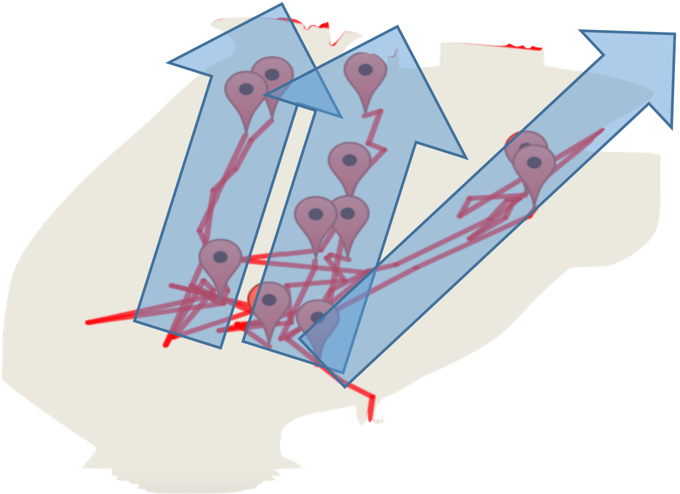
- Mechanism:
  - Unsupervised multi-level clustering of location and accelerometer data
- Identified (per animal) patterns
  - Resting
  - Grazing
  - Walking
  - Running

Unsupervised Group Pattern Learning





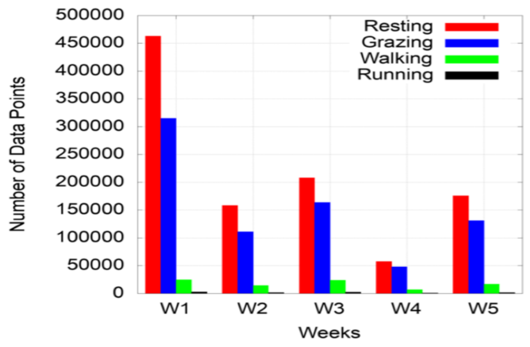
Representative of possible poacher attack



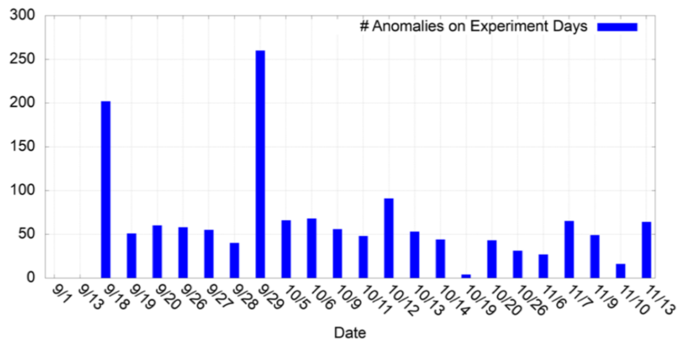
Representative of a possible predator attack

Approach: Spatiotemporal aggregation to obtain averaged group feature vectors (speed, accelerometer, direction) followed by clustering

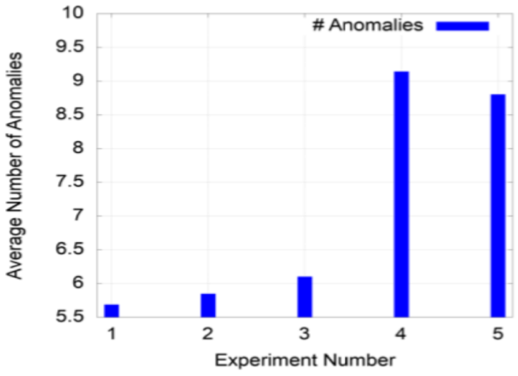
Evaluation



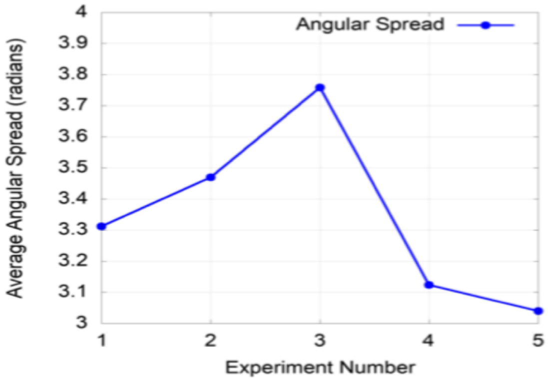
Distribution of Pattern Frequencies Over Weeks



Number of Anomalies Detected on Experiment Days



Average Number of Anomalies by Experiment Type



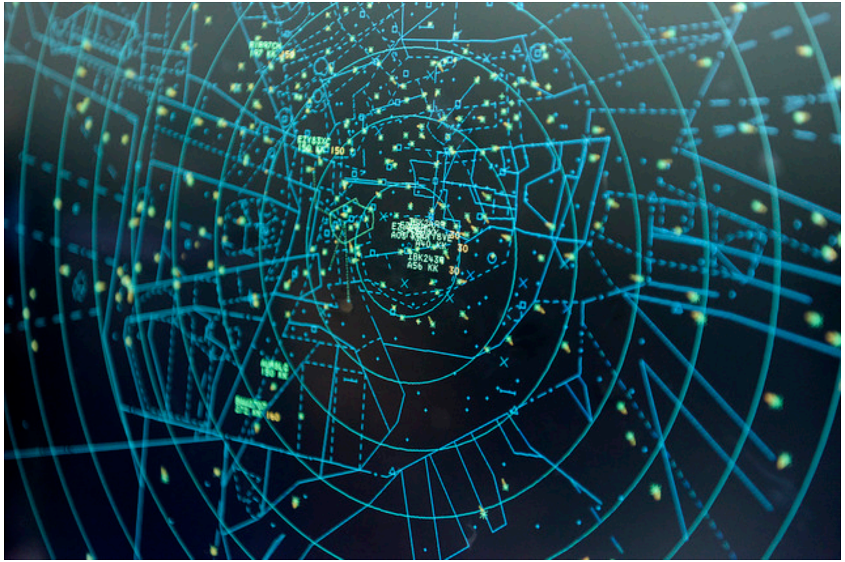
Average Group Angular Spread by Experiment Type

- Simulated experiments (5 types) were conducted over a 90 day period
- Experiment anomalies detected with **90 % accuracy**

Case Study III: Air Traffic Control

Air Traffic Control

- Sensing modality
  - GPS and RADAR
  - Typically under 20Km from Earth's surface
- Data model
  - Latitude, longitude, altitude, azimuth, ground speed, daltitude
  - Altitude is wrt mean sea level
  - Azimuth between $(0, 2\pi)$ starting with 0 = north, $\pi/2$ = east, π = south, $3\pi/2$ = west
  - Ground speed typically 0.9 mach
  - daltitude is rate of change of altitude



- Short term tracks modeled as great arcs
- Not unusual for tracks to fly over a pole (typically a point of singularity for common planar projections)

Air Traffic Control

- Automatic Dependent Surveillance Broadcast (ADS-B) is a surveillance technology in which an aircraft determines its position via satellite navigation and periodically broadcasts it, enabling it to be tracked

- Fact Sheet
  - Worldwide # flights in air
    - US day time: 9000-10000
    - US night time: 6000
  - Data gathered every minute (sometimes every 10 seconds – especially during takeoff/landing)
  - Data is neither authenticated nor encrypted and sent on a 1090 MHz channel (and thus requires RADAR based validation)

Deep Q-Learning

- Identify close approaches (encounters) between two flying objects
  - Predict encounter distance: closest distance of approach between the two flying objects
  - Predict encounter time: time at which the two flying objects are at their closest distance from each other
- Model trajectory of each flying object as a great arc/elliptic arc
  - Great arc is the shortest path between two points on a sphere
  - Unlike straight lines in Euclidean spaces, great arcs can have inflection points
- Generally a N x N problem (N: # flying objects)
  - But can be easily simplified into a m x m problem using a spatial index and altitude zones (m << N)
  - Iterative (gradient descent) algorithm to compute encounter distance/time after pruning
- ADS-B single day data for bounding box: (35, -80) to (45, -60) – roughly US North-East
  - Analysis time: one hour
  - Parallelize analysis across bounding boxes (e.g., using spatial router operator in Streams)
- Use reinforcement learning (Deep Q-learning) to provide recommendations to ATC

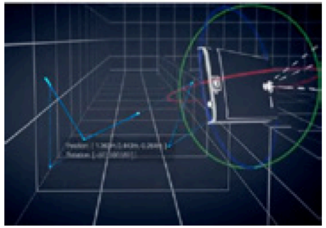
Sneak Peek into other Case Studies

AI @ Edge: 3D SLAM

Simultaneous object localization and size estimation

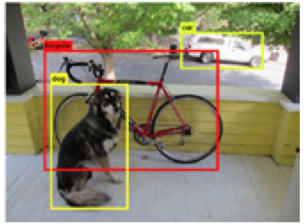
- Support for capturing point cloud data from IR and LIDAR sensors on Lenovo tango phones (Android)

Our Approach



RGB+depth cameras produce a 4 channel image which can be used for object detection and localization: distances

State-of-the-art



2D bounding boxes – no relation to actual size!

Ground Truth: 167cm
Computed: 162cm
Error: 3%

Ground Truth: 107cm
Computed: 111cm
Error: 4cms

Solution

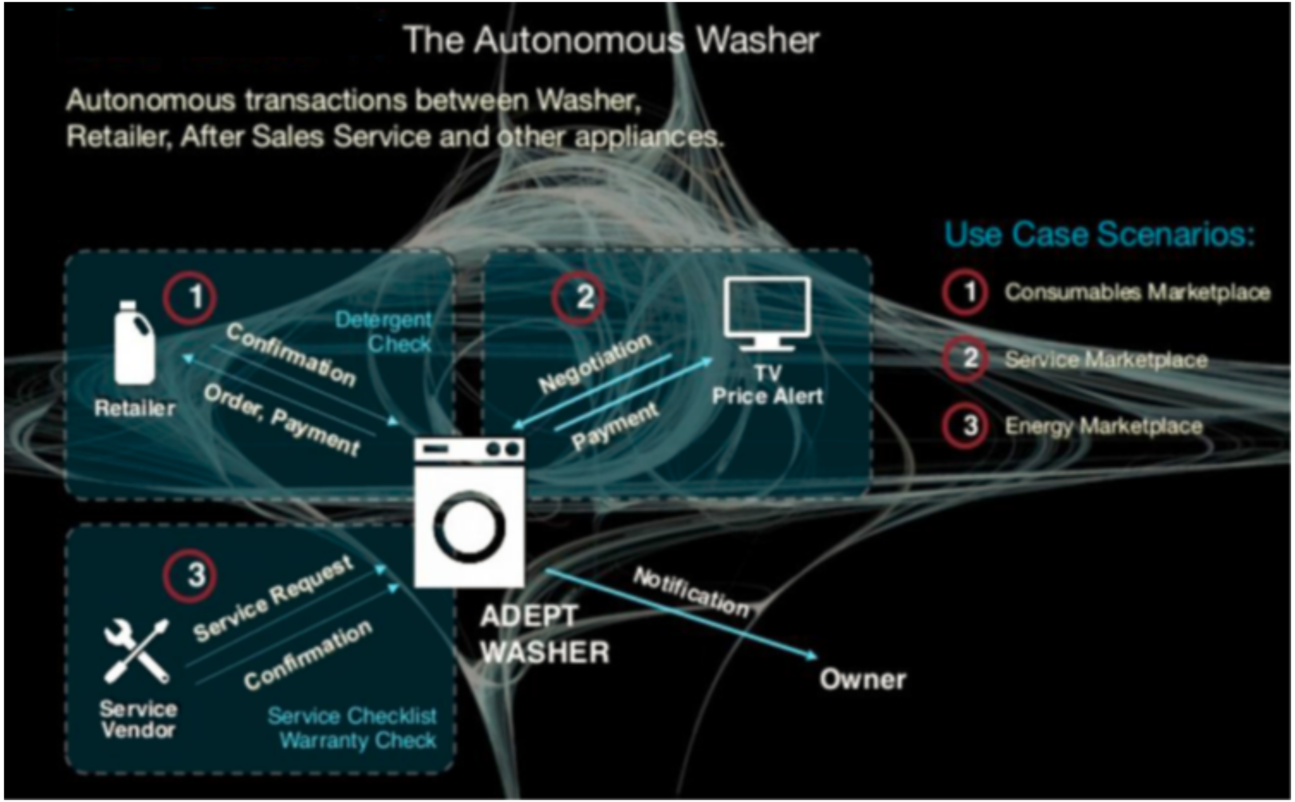
1. Depth information is captured using LIDAR or IR (LIDAR is more accurate, but more expensive)

2. Construct a four channel image (RGBD: Red-Green-Blue-Depth)

3. Train deep learning models to perform object detection and localization using real-world coordinates

4. Use a point cloud model for scoring objects (e.g., size, weight, health) if object localization is trivial (e.g., guaranteed by underlying processes used for image collection)

AI @ Edge: Autonomous Washer



ADEPT: Implementation of a decentralized blockchain based open source framework for smart devices by using Ethereum smart contracts

• Using ADEPT, an ordinary washing machine can become a semiautonomous device capable of managing its own consumables supply, performing self-service and maintenance, and even negotiating with other peer devices both in the home and outside to optimize its environment

https://www-935.ibm.com/services/multimedia/GBE03662USEN.pdf

AI @ Edge: Worker Safety

Workplace safety for remote areas with minimal infrastructure, private data

OIL RIGS

FACTORIES

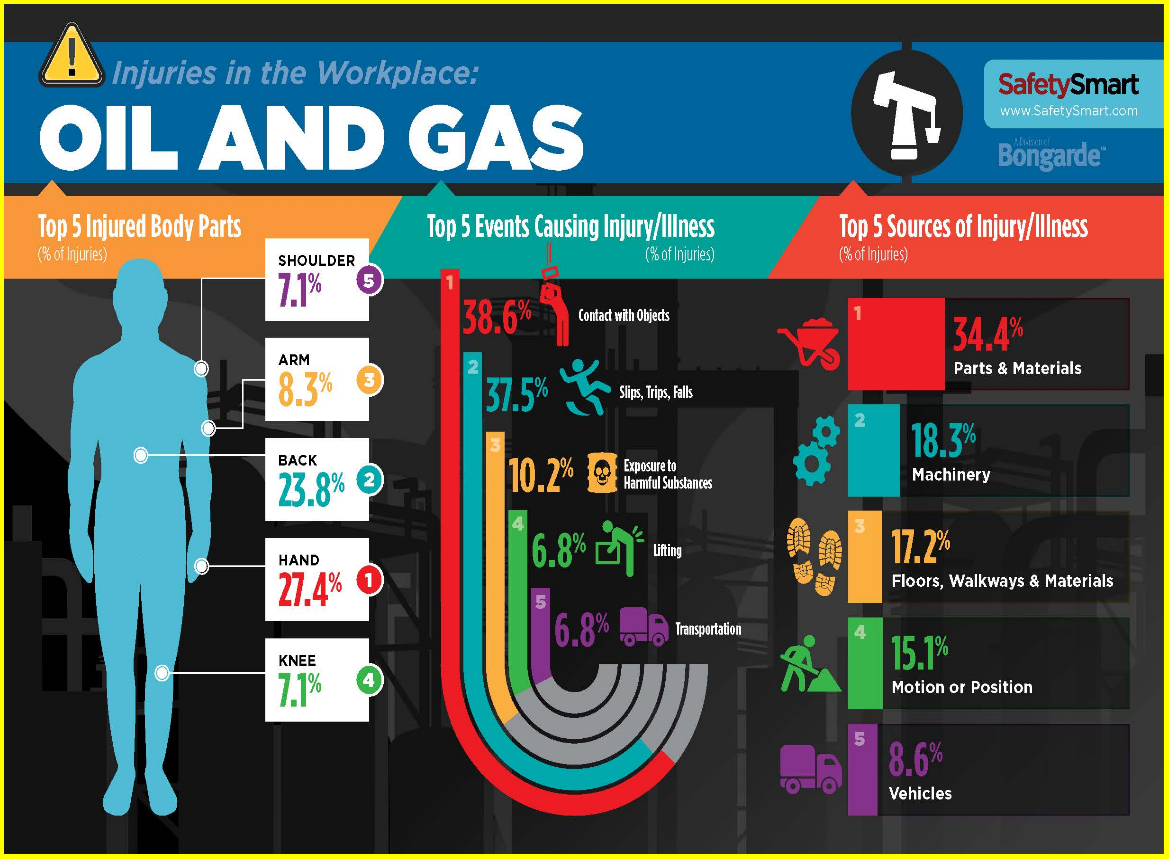
COAL MINES

CARE@HOME

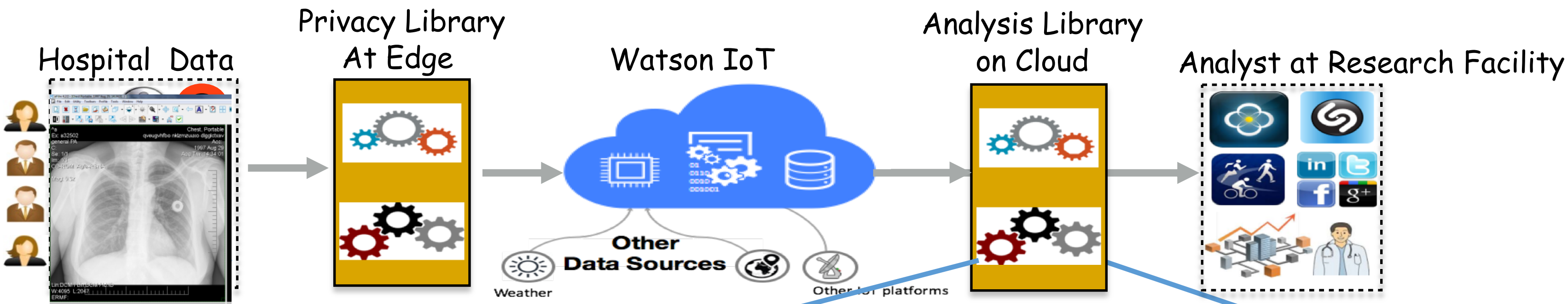
Analytics for detecting hazardous workplace conditions

Sharing safety sensors, smartphones as needed

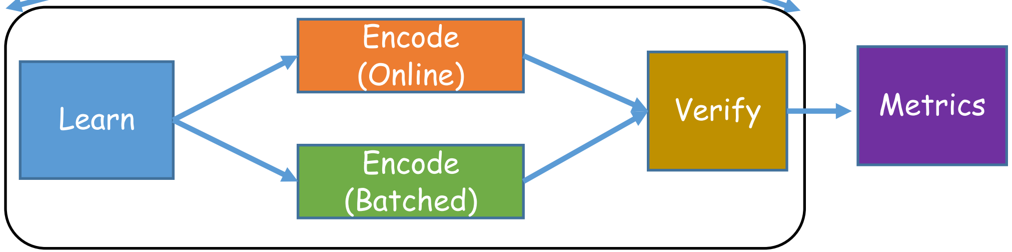
Near real-time response via co-workers, local alerting



AI @ Edge: Sensitive Healthcare Data



Privacy Analytics at Edge of hospital can obfuscate data so that analyst can get their analysis performed without seeing raw data/



"Mary Phillips is a 45-year-old woman with a history of diabetes. She arrived at New Hope Medical Center on August 5 complaining of abdominal pain. Dr. Gertrude Philippoussis diagnosed her with appendicitis and admitted her at 10 PM"

"Patient is a 42-year-old woman with a history of diabetes. She arrived at Medical Facility on August xx complaining of abdominal pain. Doctor diagnosed her with appendicitis and admitted her at yy PM."

Data captured from speech-to-text interface → anonymized → delivered via text-to-speech interface (350 ms delay)

Questions