Scientific Computing using AWS

HTCondor Week 2019

Sanjay Padhi, Ph.D

Amazon Web Services

sanpadhi@amazon.com

Amazon Web Services with HTCondor

Congratulations!!!

Development Release Series 8.7

research.cs.wisc.edu/htcondor/manual/latest/DevelopmentR

Version 8.7.8

Release Notes:

- HTCondor version 8.7.8 released on May 10, 2018.

New Features:

- *condor_annex* may now be setup in multiple regions simultaneously. Use the **-aws-region** flag with **-setup** to add new regions. Use the **-aws-region** flag with other *condor_annex* commands to choose which region to operate in. You may change the default region by setting `ANNEX_DEFAULT_AWS_REGION`. (Ticket #6632).
- Added default AMIs for all four US regions to simplify using *condor_annex* in those regions. (Ticket #6633).
- HTCondor will no longer mangle `CUDA_VISIBLE_DEVICES` or `GPU_DEVICE_ORDINAL` if those environment variables are set when it starts up. As a result, HTCondor will report GPU usage with the original device index (rather than starting over at 0). (Ticket #6584).
- When reporting `GPUsUsage`, HTCondor now also reports `GPUsMemoryUsage`. This is like `MemoryUsage`, except it is the peak amount of GPU memory used by the job. This feature only works for nVidia GPUs. (Ticket #6544).
- ...roved error messages when delegation of an X.509 proxy fails. (Ticket #6575).
- c... will no longer limit the width of the output to 80 columns when it outputs to a file or pipe. (Ticket #6643).
- Sub... of jobs via the Python bindings Submit class will now attempt to put all jobs submitted in a single transaction under ... same ClusterId. (Ticket #6649).
- ...port for *condor_schedd* option in the Python bindings. (Ticket #6619).
- ...DAP support. (Ticket #6648).

Bugs Fixe...

- ...d a problem where, when starting enough *condor_annex* instances simultaneously, some (approximately 1 in 100) instances would neither join the pool nor terminate themselves. (Ticket #6638).

ANNIVERSARY

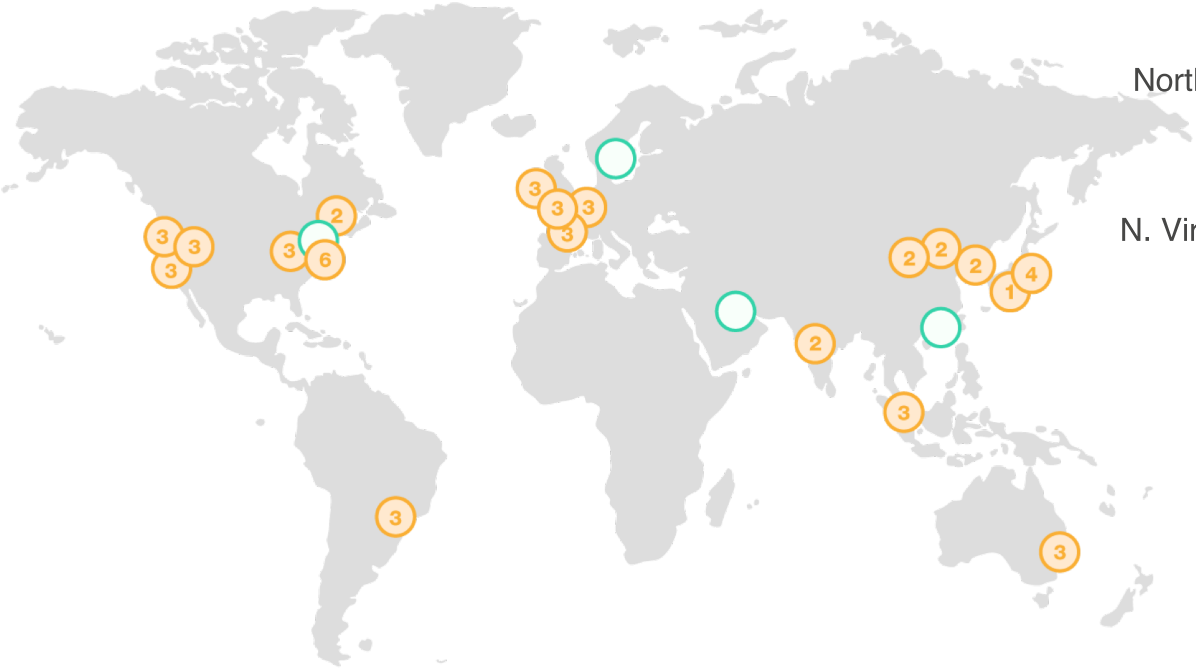
20th

aws

AWS Global Infrastructure

21 Regions – **64** Availability Zones – **155** Edge Locations

11 Regional Edge Caches in **65** cities across **29** countries



Region & Number of Availability Zones

AWS GovCloud (US)

US-East (3), US-West (3)

US West

Oregon (4)

Northern California (3)

US East

N. Virginia (6), Ohio (3)

Canada

Central (2)

South America

São Paulo (3)

Europe

Ireland (3)

Frankfurt (3)

London (3)

Paris (3), Stockholm (3)

Asia Pacific

Singapore (3)

Sydney (3), Tokyo (4),

Seoul (2), Mumbai (2)

Osaka-Local(1)

China

Beijing (2)

Ningxia (3)

New Regions

Bahrain, Cape Town, Jakarta and Milan

The AWS Secret Region is designed and built to meet the regulatory and compliance requirements of DOD, IC, etc.

Evolution in Compute Services

Virtual Server Hosting, Container management, and Serverless Computing

Amazon EC2

Provides resizable cloud-based compute capacity in the form of EC2 instances, which are equivalent to virtual servers

Amazon EC2 Container Service

A highly scalable, high performance container management service

AWS Lambda

Run code without thinking about servers.
Serverless compute for stateless code execution in response to triggers

bankinter.

airbnb NETFLIX

NASA Jet Propulsion Laboratory
California Institute of Technology

Time Inc. LAMBORGHINI

aws

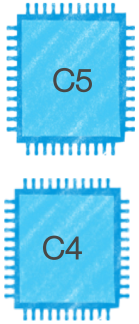
Heterogeneity in Compute Resource Instance Types & CPU

General purpose

T2.2xlarge
- 8 vCPU,
- 32GB RAM
- EBS only
- 81 cpu credit/hr

M5.24xlarge
- **96 vCPU,**
- **384GB RAM**
- **Up to 25Gps n/w**
- **EBS only**
- **9k EBS Mbps**
- **New Nitro light hypervisor + dedicated h/w**

Compute optimized



C5.18xlarge
- **72 vCPU,**
- **144GB RAM**
- **EBS only**
- **9k EBS Mbps**
- **Up to 25 Gbps w/ENA**

Storage and IO optimized

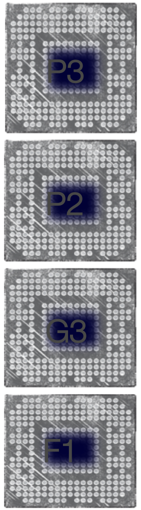
H1.16xlarge
- 64 vCPU
- 256GB RAM
- 8 x 2TB HDD
- 25 Gbps

D2.8xlarge
- 36 vCPU
- 256GB RAM
- 8 x 2TB HDD
- 25 Gbps

I3.16xlarge
- 64 vCPU
- 488GB RAM
- 8 x 2TB NVMe SDD
- 25 Gbps

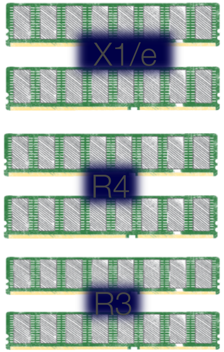
I3.metal (preview)
- 36 cores/72
- 512GB RAM
- 8 x 2TB NVMe SDD
- 25 Gbps

Accelerated Computing



P3.16xlarge
- **8 GPU Tesla V100**
- **5k CUDA/640 Tensor cores**
- **488GB RAM**
- **64GB GPU RAM**
- **NVLink p-2-p**

Memory optimized



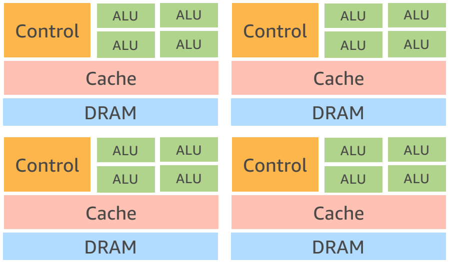
X1e.32xlarge
- 128 vCPU,
- 4TB RAM
- 2 x 1.9TB SSD
- 14k EBS Mbps

R4.16xlarge
- 64 vCPU,
- 488GB RAM
- SSD EBS
- 25 Gbps

Selection of different Intel Xeon processors
- 2.3/2.4 GHz Intel Broadwell/Haswell CPUs: M4, I3, H1, D2, G3, P3/2 instance types
- 2.9 GHz Intel Haswell CPUs: C4
- 2.5 GHz Intel Platinum CPUs: w/AVX-512 instruction set: M5
- 3.0 GHz Intel Platinum CPUs: w/AVX-512 instruction &set Turbo up to 3.5Ghz: C5

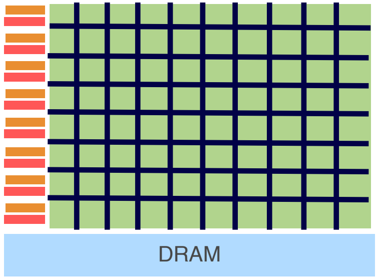
CPUs vs GPUs vs FPGA for Compute

CPU



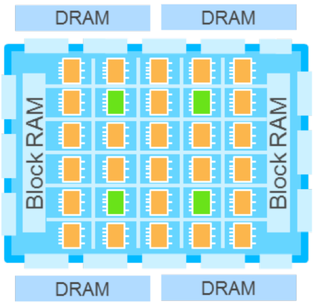
- 10s-100s of processing cores
- Pre-defined instruction set & datapath widths
- Optimized for general-purpose computing

GPU



- 1,000s of processing cores
- Pre-defined instruction set and data path widths
- Highly effective at parallel execution

FPGA

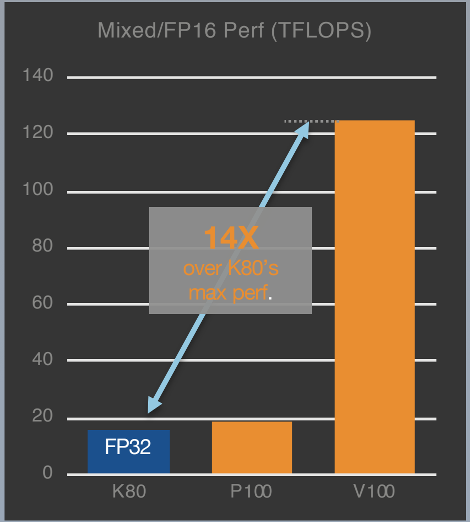


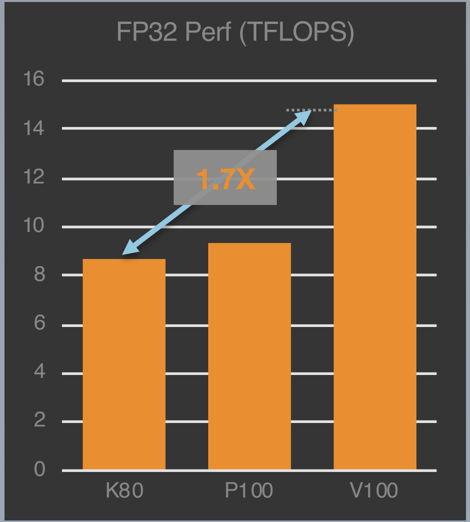
- Millions of programmable digital logic cells
- No predefined instruction set or datapath widths
- Hardware timed execution

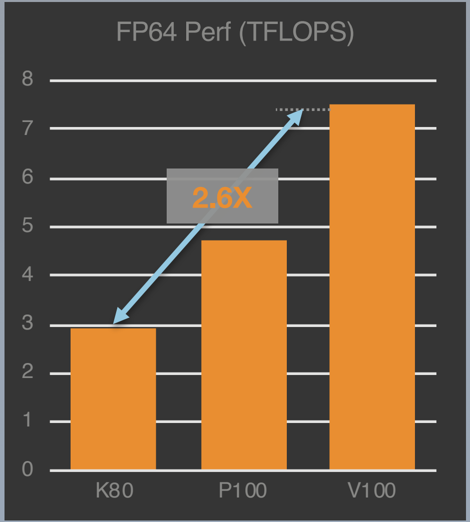
GPU Performance Comparison

ResNet-50 Training Performance
(Using Synthetic Data, TensorFlow 1.5)

- P2 Instances use K80 Accelerator (Kepler Architecture)
- P3 Instances use V100 Accelerator (Volta Architecture)







Security Requirements

ISO 9001
Global Quality Standard

ISO 27001
Security Management Standard

ISO 27017
Cloud Specific Controls

ISO 27018
Personal Data Protection

PCI DSS Level 1
Payment Card Standards

SOC 1
Audit Controls Report

SOC 2
Compliance Controls Report

SOC 3
General Controls Report

CJIS
Criminal Justice Information Services

DoD SRG
DoD Data Processing

FDA
Food and Drug Administration

FedRAMP
Government Data Standards

FedRAMP TIC
FedRAMP-Trusted Internet Connection

FERPA
Educational Privacy Act

FIPS
Government Security Standards

FISMA
Federal Information Security Management

GXP
Quality Guidelines and Regulations

HIPAA
Protected Health Information

SEC Rule 17a-4(f)
Financial Data Standards

ITAR
International Arms Regulations

FISC [Japan]
Financial Industry Information Systems

IRAP [Australia]
Australian Security Standards

MLPS Level 3 [China]
Multi-Level Protection

MTCS Tier 3 [Singapore]
Multi-Tier Cloud Security Standard

My Number Act [Japan]
Personal Information Protection

DNB [Netherlands]
Dutch Financial Regulations

EU Data Protection
Data Protection Framework

G-Cloud [UK]
UK Government Standards

IT-Grundschutz [Germany]
Baseline Protection Methodology

Privacy Shield
EU-US Data Transfer

UK Cyber Essentials Plus
Cyber Threat Protection

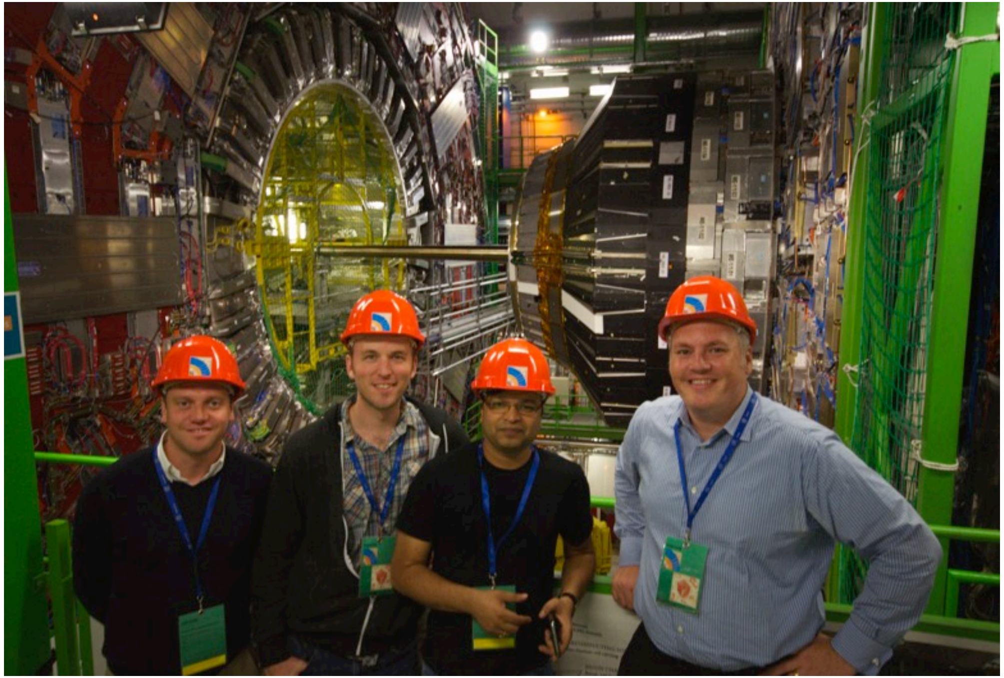
https://aws.amazon.com/compliance/

Predictive Analytics in Scientific Computing

Large Hadron Collider

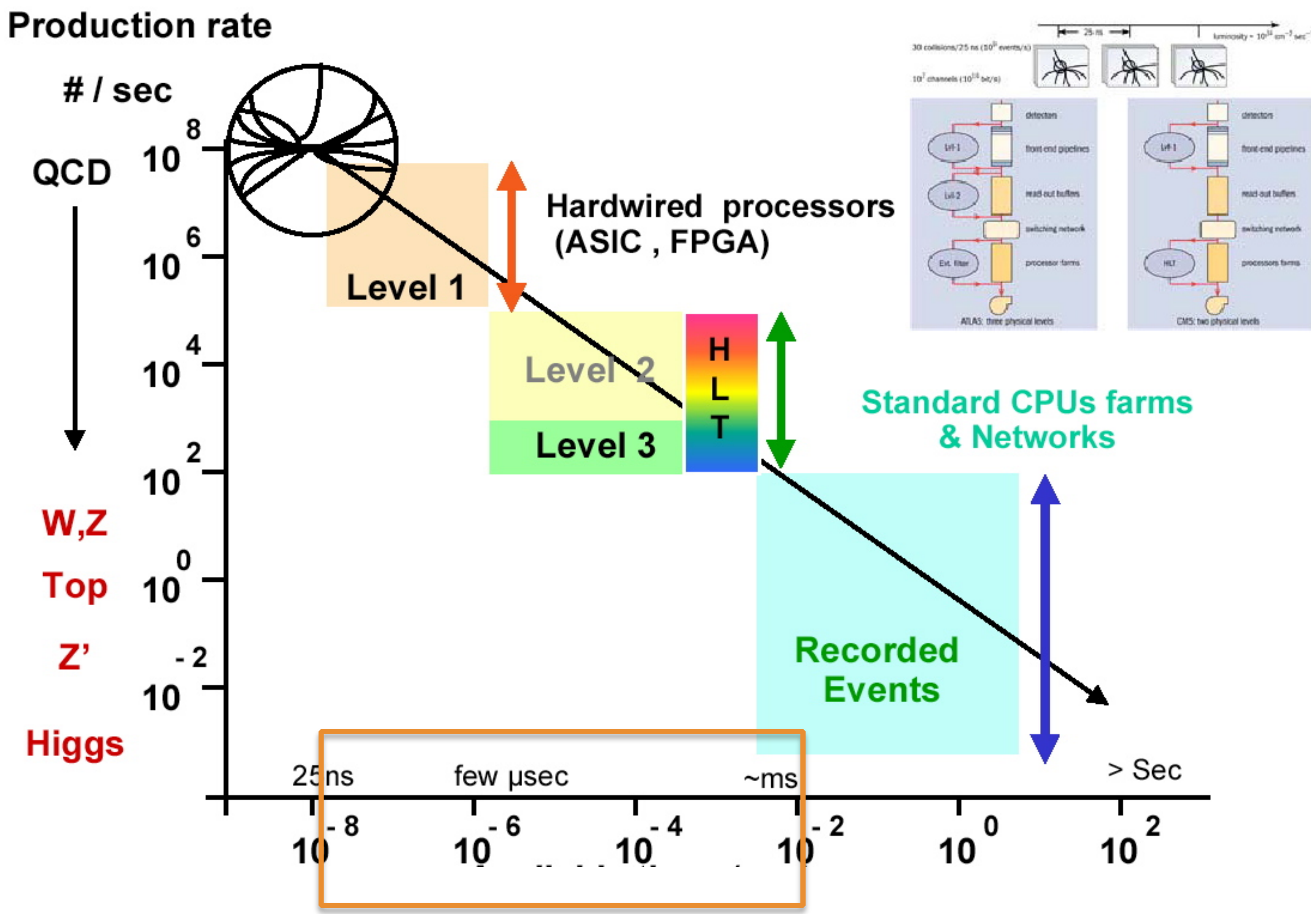
The Large Hadron Collider @ CERN includes 6,000+ researchers from over 40 countries and produces approximately 25PB of data each year.

The ATLAS and CMS experiments are using AWS for Monte Carlo simulations, processing, and analysis of LHC data.

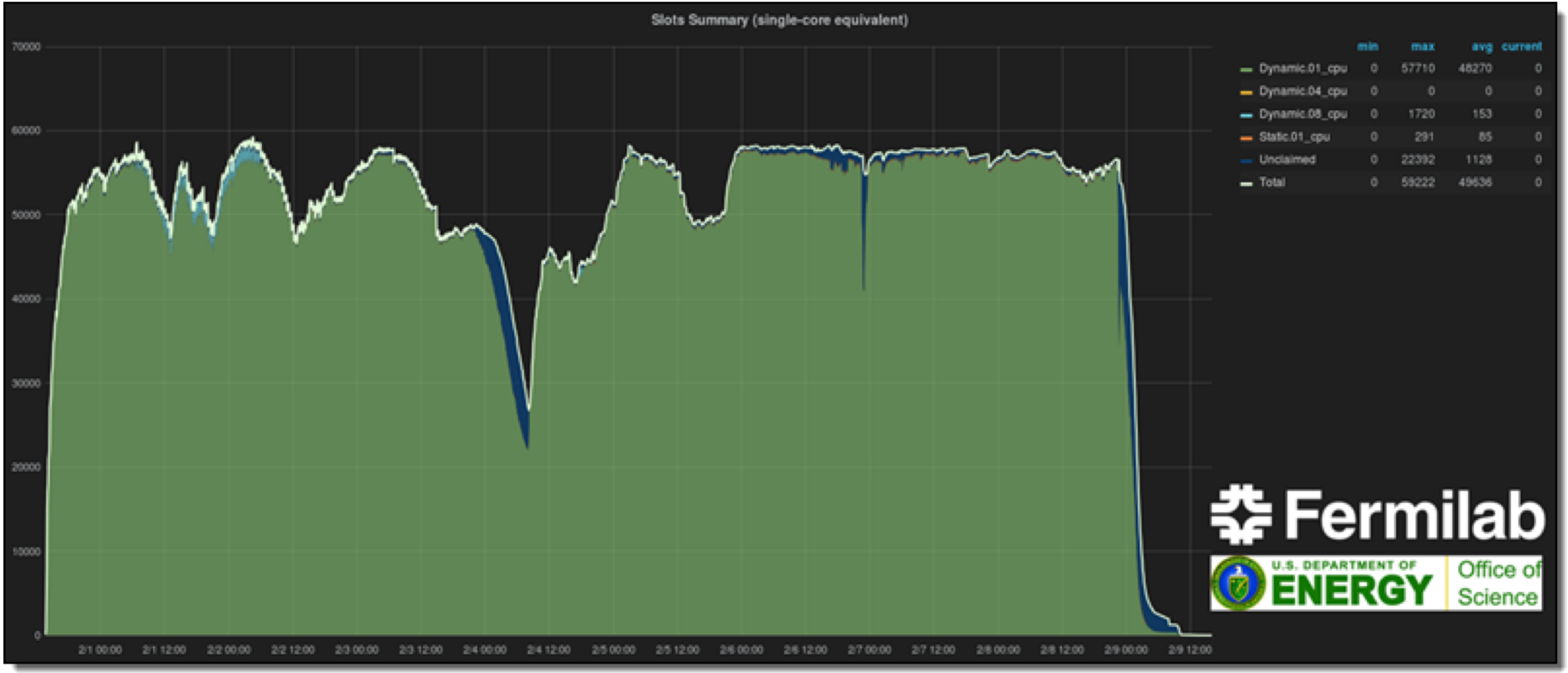


aws

Analytics at the LHC



Elasticity in Computing: On demand auto-expansion to AWS – HTCondor based

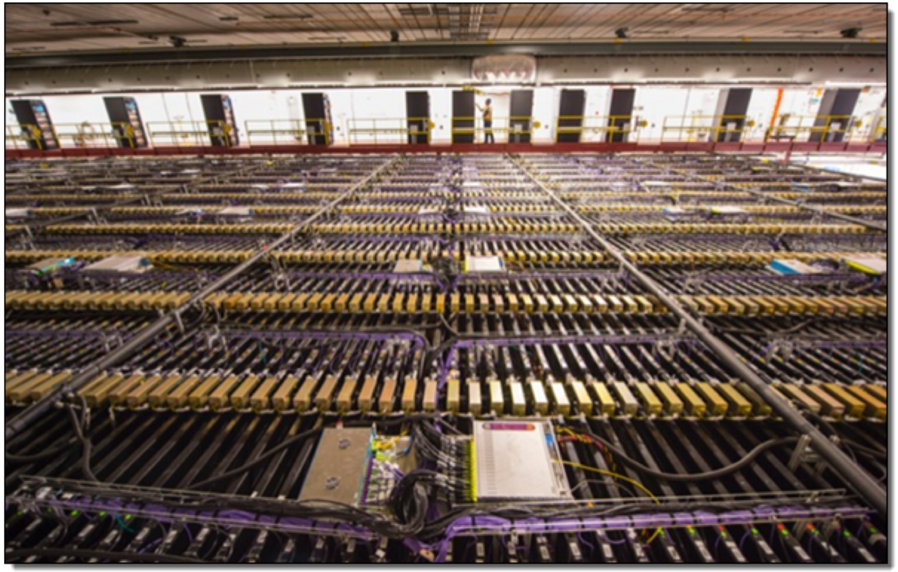


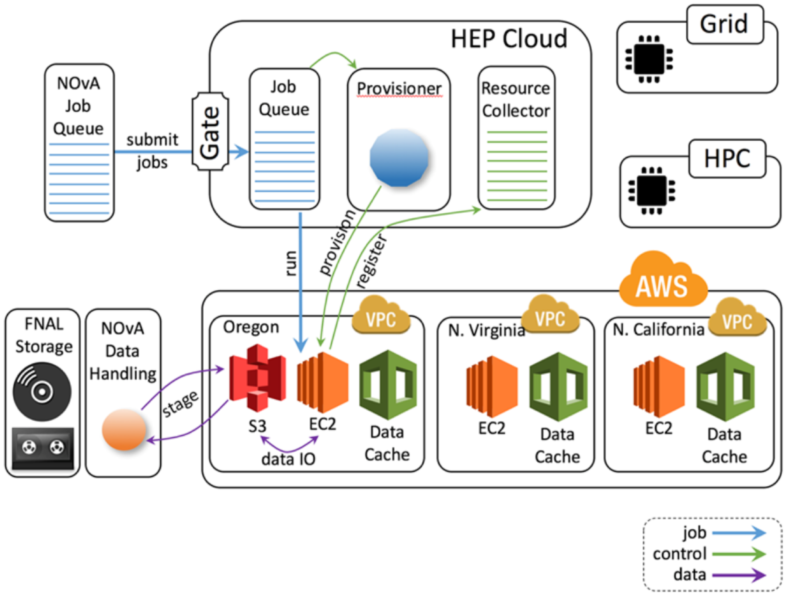
~60,000 slots using AWS spot instances. A factor of 5 larger than Fermilab capacity!

https://aws.amazon.com/blogs/aws/experiment-that-discovered-the-higgs-boson-uses-aws-to-probe-nature/

aws

NOvA uses AWS to Shed Light on Neutrino Mysteries



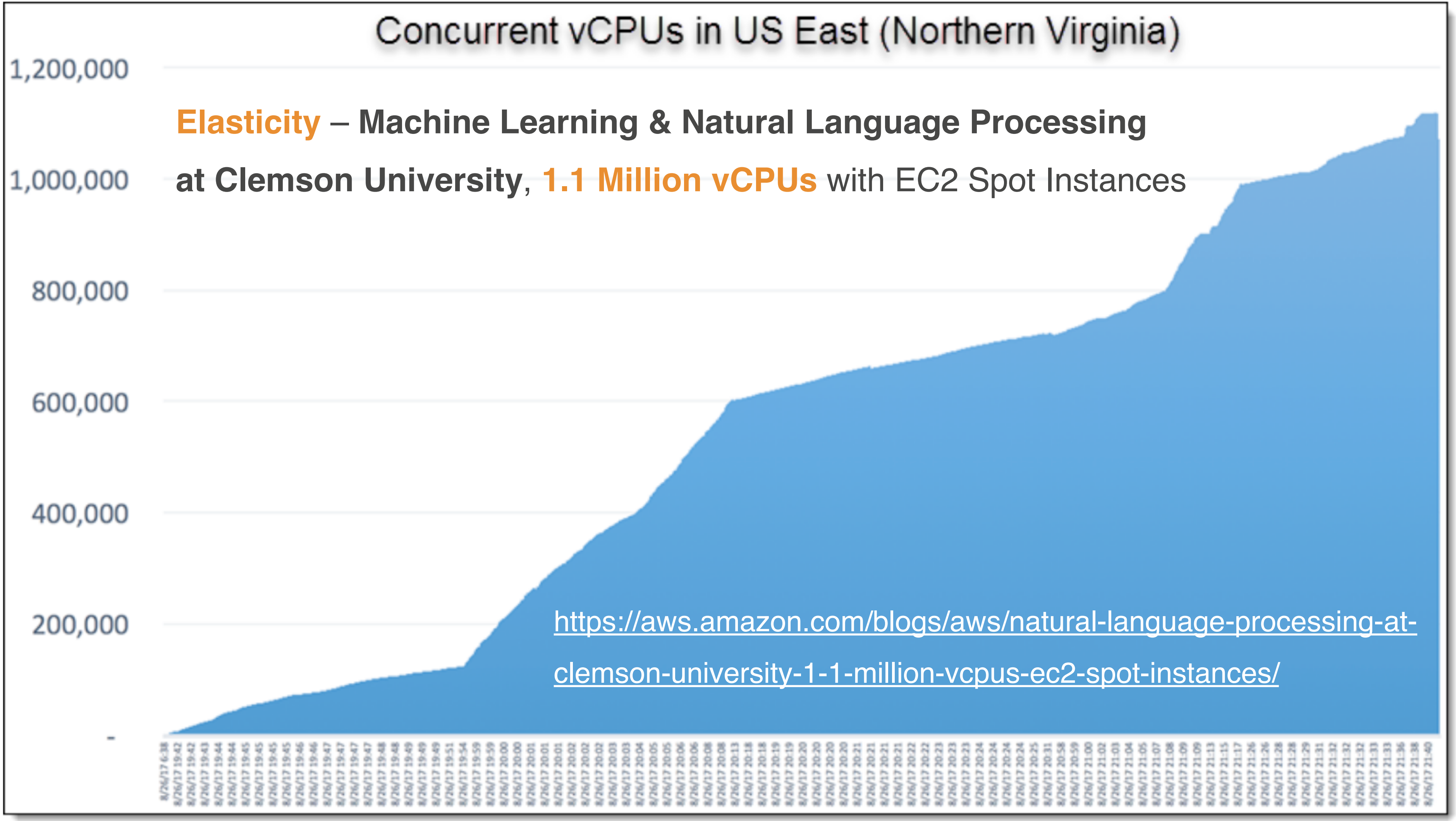


[Peter Shanahan]() (Co-spokesperson of the NOvA experiment):
"Our experience with Amazon Web Services shows its potential as a reliable way to meet our peak data processing needs at times of high demand"

https://aws.amazon.com/blogs/aws/nova-uses-aws-to-shed-light-on-neutrino-mysteries/

Neutrinos are ghost like particles → Needed advanced ML analytics to detect

Elasticity – **Machine Learning & Natural Language Processing at Clemson University**, **1.1 Million vCPUs** with EC2 Spot Instances



https://aws.amazon.com/blogs/aws/natural-language-processing-at-clemson-university-1-1-million-vcpus-ec2-spot-instances/

Available in AWS Marketplace

HTCondor High Throughput Computing

HTCondor 8.7.2 on Amazon Linux 2017.03 (GPU-ready)

Sold by: Center for High Throughput Computing   Latest Version: v8.7.2

The HTCondor high-throughput computing system, version 8.7.2, installed on Amazon Linux 2017.03.

Linux/Unix   ☆☆☆☆☆ (0)   Free Tier

Continue to Subscribe

Save to List

Typical Total Price
$0.100/hr
Total pricing per instance for services hosted on m4.large in US East (N. Virginia). View Details

Overview   Pricing   Usage   Support   Reviews

Product Overview

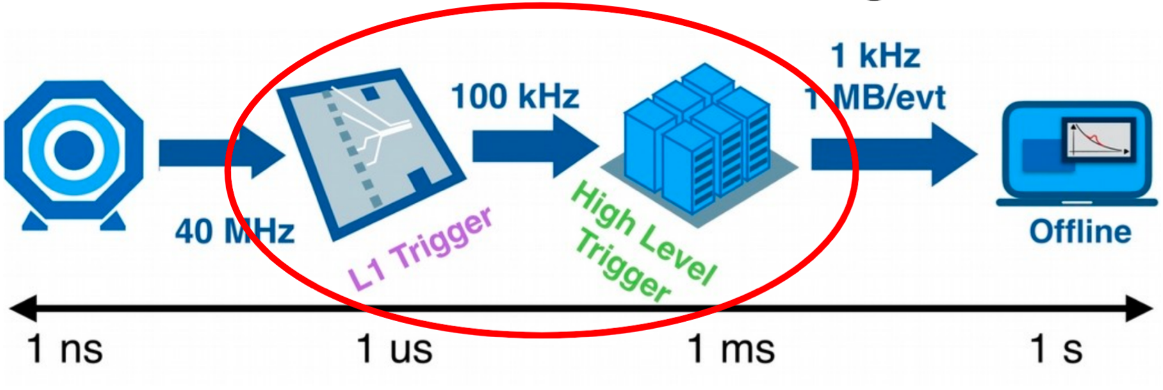
The HTCondor high-throughput computing system is a workload management system for compute-intensive jobs.

Version	v8.7.2
Sold by	Center for High Throughput Computing
Categories	High Performance Computing
Operating System	Linux/Unix, Amazon Linux 2017.03
Fulfillment Methods	Amazon Machine Image

Highlights

- Complete single-node HTCondor pool, ready to run
- Use the condor_annex command to easily add more instances to this HTCondor pool

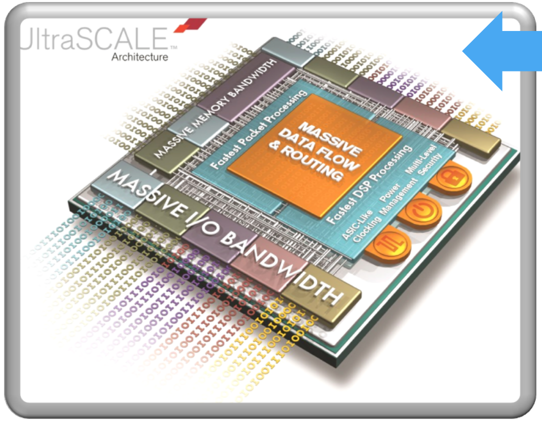
LHC Data Processing



- DNNs have the potential to greatly improve physics performance in the trigger system
- In order to implement an algorithm, need to ensure inference latencies of μs (ms) for L1 (HLT)
  - For L1, this means we *must* use FPGAs
- ***How can we run neural network inference quickly on an FPGA?***

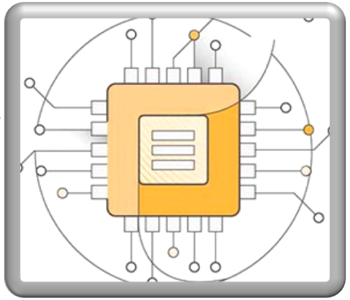
FPGA Acceleration

FPGA handles compute-intensive, deeply pipelined, hardware-accelerated operations



Application

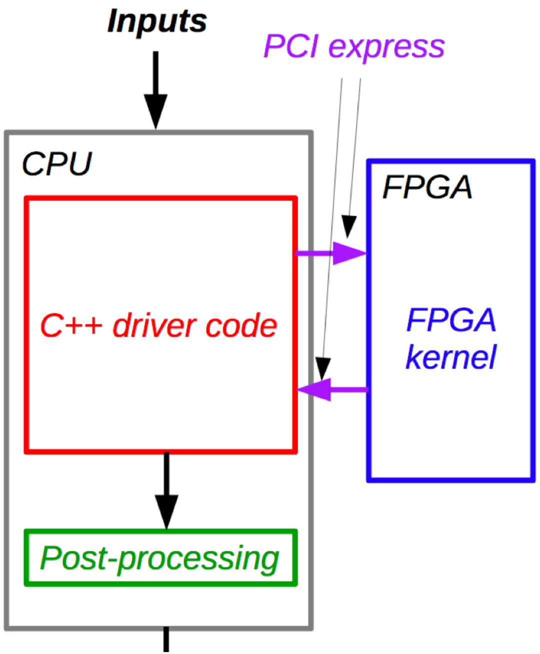
```
....
....
....
....
....
....
....
module filter1 (clock, rst, strm_in, strm_out)
  integer i,j; //index for loops
  always@(posedge clock)
  for (i=0; i<NUMUNITS; i=i+1)
      tmp_kernel[j] = k[i*OFFSETX];
....
....
```



CPU handles the rest

Acceleration with AWS

- Development for FPGA kernel and CPU host code is done with SDAccel environment
  - Invokes Vivado HLS under the hood, produces traditional synthesis reports etc.
- Run host code on CPU, manages data transfer and FPGA kernel execution
- **`hls4ml`** project only needs to be wrapped to provide specific inputs/outputs for SDAccel to interface properly
  - Can be done generically
  - Have accelerated variety of hls4ml projects on AWS F1
- Limited in speed by I/O bandwidth



An Acceleration Case Study (2)

- Have successfully implemented/run the network inference on AWS using hls4ml/SDAccel
- https://github.com/drankincms/AccelFPGA
- Including data transfer to/from CPU, whole *FPGA inference process takes **2 ms*** for all 16k HCAL channels
  - Inference alone takes 80 us (70 ns for one inference)
- Has been tested inside standard CMS software code environment, using high-level trigger job
  - Every event sends input features to FPGA, waits for callback
- *Iterative fit procedure takes **50 ms*** for same inputs
- FPGA inference is a fixed-latency procedure, iterative fit is not
- Inference on CPU or GPU also significantly faster than iterative fit
  - FPGA inference fastest

Algorithm	***Architecture***	***Time/event (ms)***
Iterative fit	***CPU***	**50**
NN Inference	***CPU***	**15**
NN Inference	***GPU***	**12**
NN Inference	***FPGA***	**2**

A key component in autonomous vehicle and fast triggering systems, learn how FPGAs do real-time DNN inference in this hands-on course. Topics include:

- Model compression and quantization
- High-level synthesis
- Firmware implementation
- Model acceleration on cloud FPGAs

The class is given by **Dr. Jennifer Ngadiuba (CERN)** and **Dr. Dylan Rankin (MIT)** and consists of half a day of lectures as well as a hands-on sessions.

You'll learn how to compress and synthesise your own TensorFlow model, as well as implement it on a Xilinx FPGA on the Amazon cloud.

The course is targeted at PhD, Postdocs and Professors, but others will be allowed to participate if there are available places.

The lectures and hands-on session will take place at the UZH Irchel Campus in the Physik Institut (building 36)

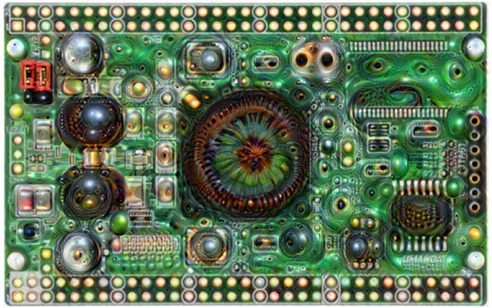
https://indico.cern.ch/event/769727/

Universität Zürich[UZH]

Physik-Institut

How to do ultrafast Deep Neural Network inference on FPGAs

6. February 2019
Physik Institut - Universität Zürich



A key component in autonomous vehicle and low-latency triggering systems, learn how FPGAs do real-time DNN inference in this hands-on course. Topics include:

- Model compression and quantization
- High-level synthesis
- Firmware implementation
- Model acceleration on cloud FPGAs

Lecturers:
Dr. Jennifer Ngadiuba (CERN)
Dr. Dylan Rankin (MIT)

Registration and further info at
indico.cern.ch/e/FPGA4HEP

Organizers:
Thea Årrestad (UZH)
Jennifer Ngadiuba (CERN)
Dylan Rankin (MIT)
Maurizio Pierini (CERN)
Ben Kilminster (UZH)

Organizers:
Thea Aarrestad (UZH)
Jennifer Ngadiuba (CERN)
Dylan Rankin (MIT)
Maurizio Pierini (CERN)
Ben Kilminster (UZH)

All course material can be found as attachments to the timetable, or at

https://github.com/FPGA4HEP/course_material

Methods to use FPGA (AWS F1 Instance)

1 Hardware Engineers/Developers

- Developers who are comfortable programming FPGA
- Use F1 Hardware Development Kit (HDK) to develop and deploying custom FPGA accelerations using Verilog and VHDL

2 Software Engineers/Developers

- Developers who are not proficient in FPGA design
- Use OpenCL to create custom accelerations

3 Software Engineers/Developers

- Developers who are not proficient in FPGA design
- Use pre-build and ready to use accelerations available in AWS Marketplace

Children's Hospital of Philadelphia And Edico Genome Achieve Fastest-Ever Analysis Of 1,000 Genomes

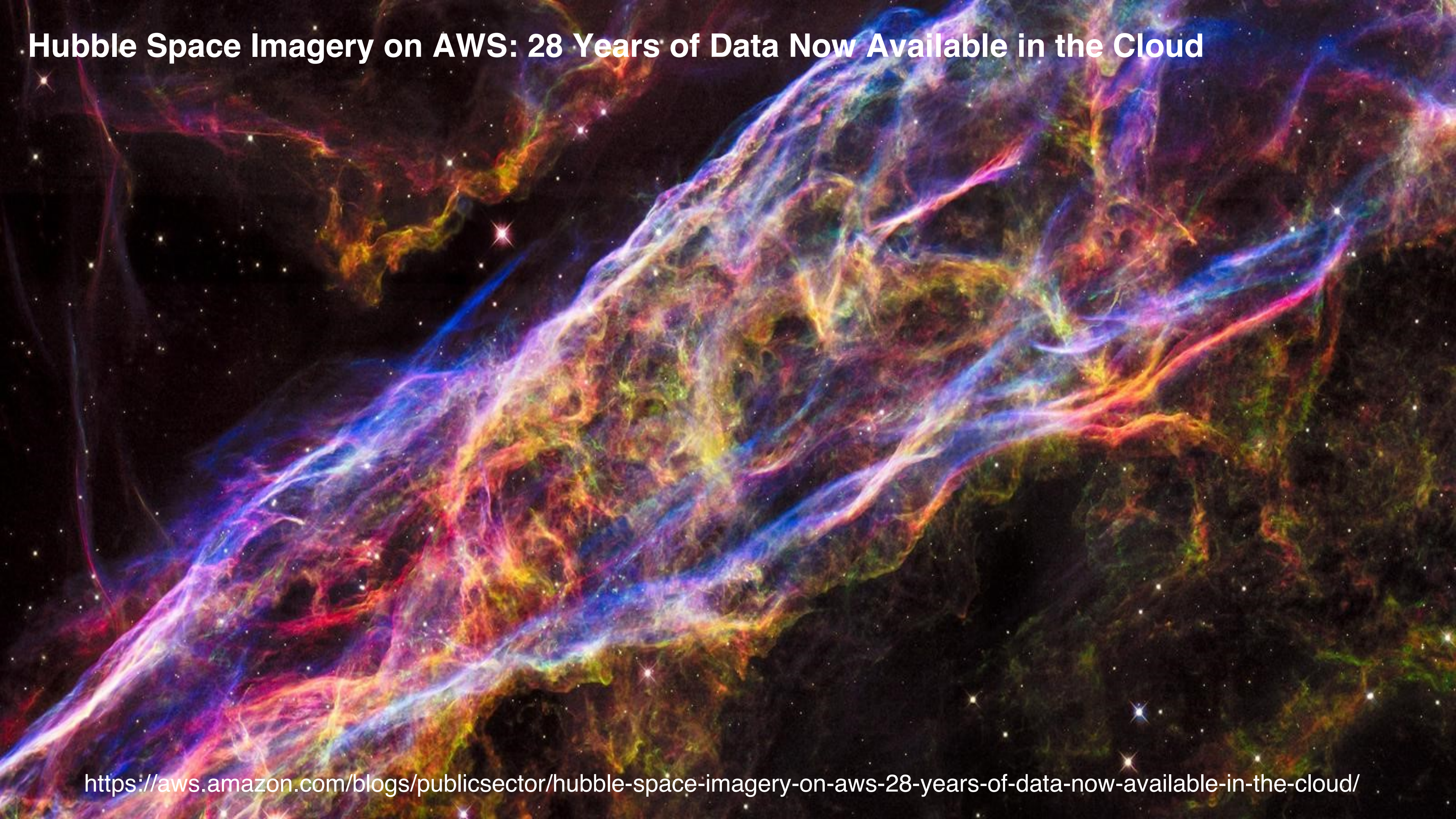
GUINNESS WORLD RECORDS title for *Fastest time to analyze 1,000 human genomes*

The **Amazon EC2 F1 instances**, with Xilinx Virtex UltraScale+ field programmable gate arrays (FPGAs) was used for 1,000 diverse pediatric genomes.

The study was completed in two hours and twenty-five minutes.

https://www.prnewswire.com/news-releases/childrens-hospital-of-philadelphia-and-edico-genome-achieve-fastest-ever-analysis-of-1000-genomes-300540026.html

Hubble Space Imagery on AWS: 28 Years of Data Now Available in the Cloud



https://aws.amazon.com/blogs/publicsector/hubble-space-imagery-on-aws-28-years-of-data-now-available-in-the-cloud/

Research Collaborations

Research Collaboration: AMPLab

AMP stands for "Algorithms, Machines, and People"

Research Outcomes

http://mesos.apache.org

http://www.mlbase.org

http://spark.apache.org

http://spark.apache.org

https://databricks.com

http://snap.cs.berkeley.edu

Research Collaboration: RISELab (Real-time Intelligent Secure Execution)

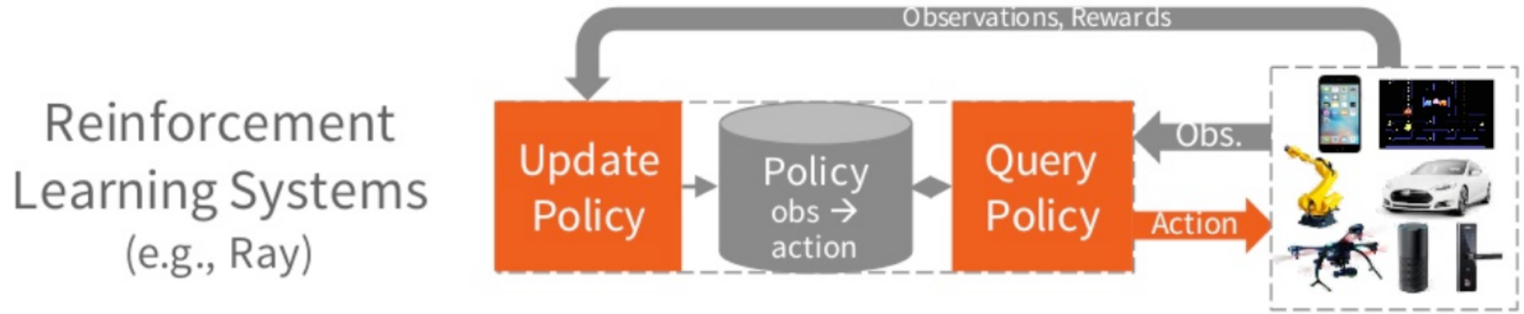
Collaborative 5-year effort between UC Berkeley, National Science Foundation and Industry

AWS as a founding partner - https://rise.cs.berkeley.edu/

Goal: Develop open source platforms, tools and algorithms for intelligent real-time decisions on live-data

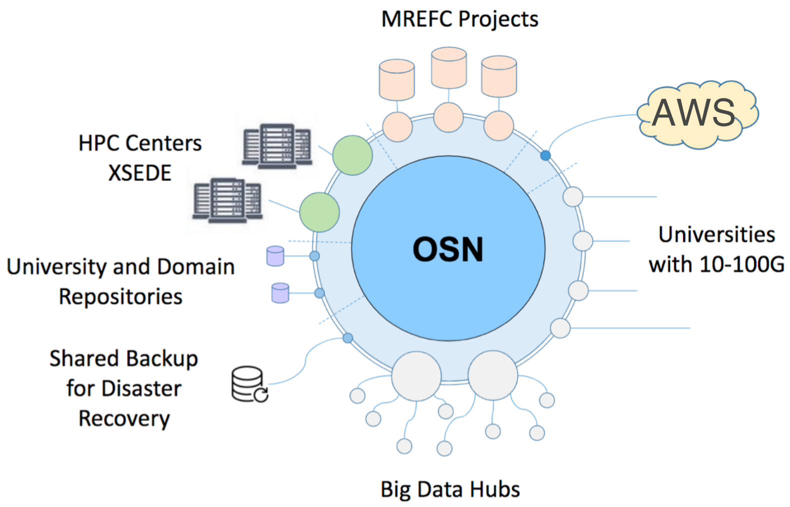
- Researchers at RISELab use AWS to rapidly prototype and develop systems at scale
- Resulted in Apache Spark, developed on AWS and integrated with core services

From live data to real-time decisions



Research Data Management

AWS part of Nationwide Open Storage Network (Funded by NSF)



NSF supports development of new nationwide data storage network

The Open Storage Network will enable researchers to manage data more efficiently than ever before

Alex Szalay is shown in the Data-Scope, a resource for computationally-intensive computing at JHU.

Credit and Larger Version

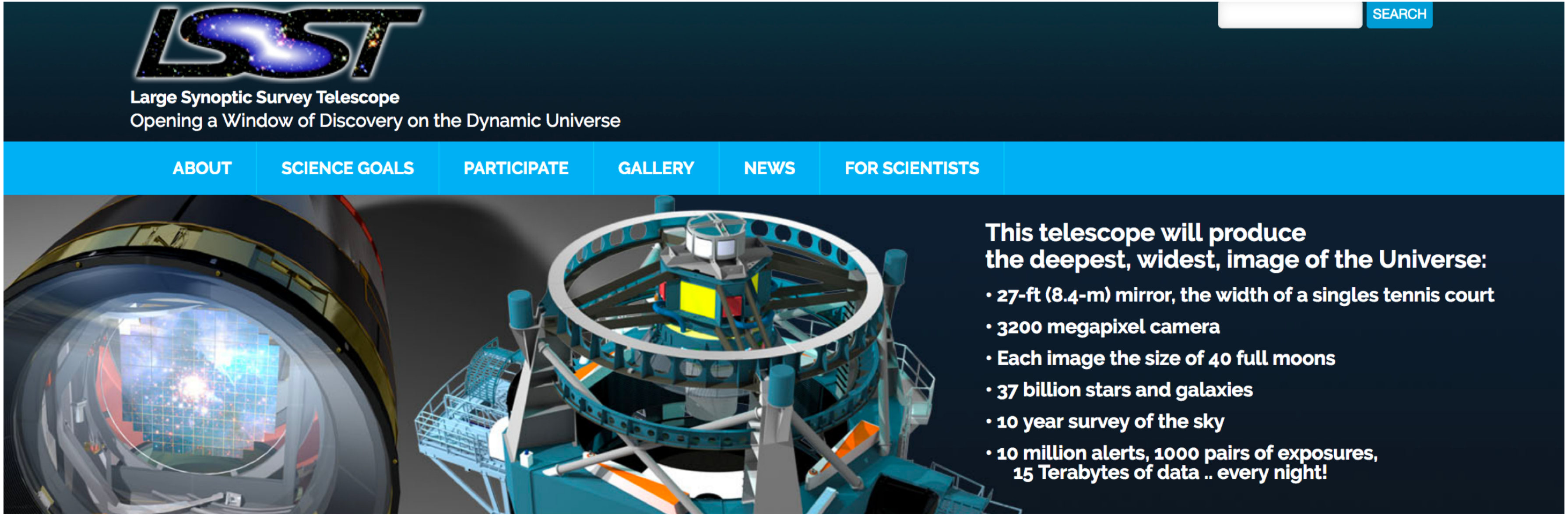
June 7, 2018

The National Science Foundation (NSF) is announcing a $1.8 million grant for the initial development of a data storage network over the next two years. A collaborative team will combine their expertise, facilities and research challenges to develop the Open Storage Network (OSN). OSN will enable academic researchers across the nation to work with and share their data more efficiently than ever before.

Collaboration includes:

- JHU, UCSD, MGHPCC, AWS, Globus, Internet2, 4 NSF Big Data Hub, etc

Study: Processing LSST using AWS



LARGE SYNOPTIC SURVEY TELESCOPE

LSST + Amazon Web Services Proof of Concept

Kian-Tat Lim, Leanne Guy, and Hsin-Fang Chiang

2019-03-13

Forbes

FEB 18, 2018 @ 02:42 PM 1,933

NSF's New Initiative To Bring The Cloud Era To Academic Big Data Research

NSF Partners With 3 Cloud Providers for Data Science Research Support Program

The National Science Foundation has partnered with Amazon Web Services, Microsoft and Google to support research projects in the data science and engineering field. NSF said Wednesday it will obligate nearly $30 million to the Critical Techniques, Technologies and Methodologies for Advancing Foundations and Applications of Big Data Sciences and Engineering program...

Read More

NSF Adds $30M to BIGDATA Program; AWS, Google, and Azure Participate

By Staff

Collaborative programs with the National Science Foundation (NSF)

- 2017/2018: The NSF Big Data program supported by multiple directorates at NSF
- 2019: Collaboration with NSF CISE/OAC on Campus Cyberinfrastructure
- 2019 : NSF CISE - Exploring Cloud for Acceleration of Science (E-CAS)
- 2019: NSF Cloud Access Model - All

Examples of Research Collaboration with NSF (2018)

- *Automating Analysis and Feedback to Improve Mathematics Teachers' Classroom*

  University of Colorado, Boulder
- *Collaborative Research: Protecting Yourself from Wildfire Smoke: Big Data Driven Adaptive Air Quality Prediction Methodologies*

  University of Nevada, Reno
- *Collaborative Research: TIMES: A tensor factorization platform for spatio-temporal data*

  Emory University
- *Collaborative Research: Optimizing Log-Structured-Merge-Based Big Data Management Systems*

  University of California, Riverside
- *Collaborative Research: Intelligent Solutions for Navigating Big Data from the Arctic and Antarctic*

  Texas A&M University Corpus Christi

"This NSF big data award, coupled with AWS's advanced computational and analytic services, is expected to help unlock the secrets of interactions among biomolecules that drive human and animal biological processes."

Dr. Bin Yu, Chancellor's Professor at University of California,

Examples of Research Collaboration with NSF (2017)

- *Detecting Financial Market Manipulation: An Integrated Data- and Model-Driven Approach*

  University of Michigan, Georgia Tech
- *Scalable and Interpretable machine learning: bridging mechanistic and data-driven modeling in the biological sciences*

  University of California, Berkeley
- *Taming Big Networks via Embedding*

  University of Virginia, University of Illinois at Urbana-Champaign
- *Domain Adaptation Approaches for Classifying Crisis Related Data on Social Media*

  Kansas State University, University of North Texas and Pennsylvania State University
- *Distributed Semi-Supervised Training of Deep Models and Its Applications in Video Understanding*

  University of Central Florida

"In today's era of data-driven science and engineering, we are pleased to work with the AWS Research Initiative via the NSF BIGDATA program, to provide cloud resources for our Nation's researchers to foster and accelerate discovery and innovation."

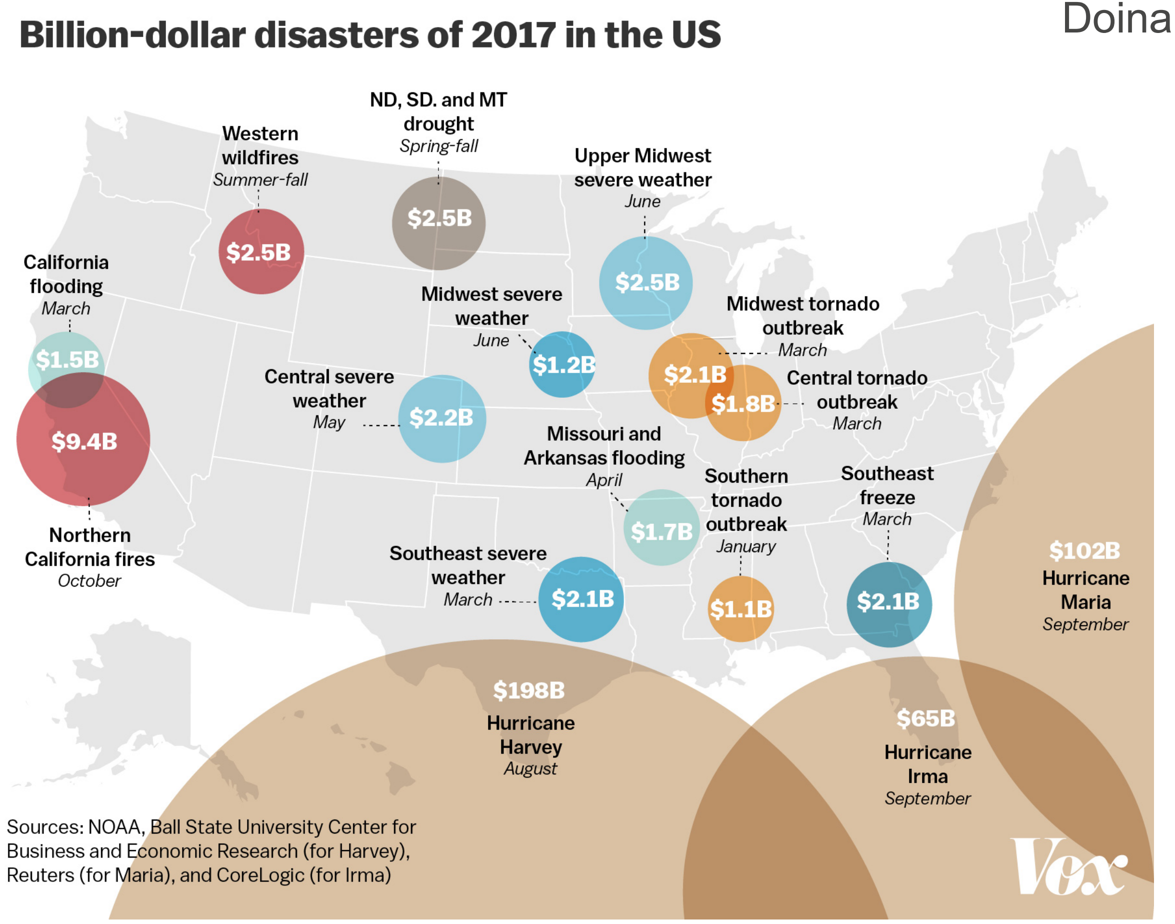
Dr. Jim Kurose, Assistant Director, CISE, National Science Foundation (NSF)

Natural Language Processing is part of Artificial Intelligence (AI):

- Intersects with Computers and human (natural) languages
- Involves Speech recognition, natural language understanding & natural language generation

Domain Adaptation Approaches for Classifying Crisis Related Data on Social Media

Doina Caragea, KSU



Methodology

Data Collection

JSON tweets

Data Extraction

Tweet id, create time, text

Data Processing

Stop words, special characters, URLs, Emails

Topics Modeling

Streaming Corpus

Latent Dirichlet Allocation

Analysis

Preparedness, During Hurricane, Aftermath

Hurricane timeline

Classes of machine learning algorithms

Supervised learning [Imran et al., 2013; Ashktorab et al., 2014; Caragea et al., 2014; Imran et al., 2018]

- Labeled tweets needed, but not readily available for an emergent disaster

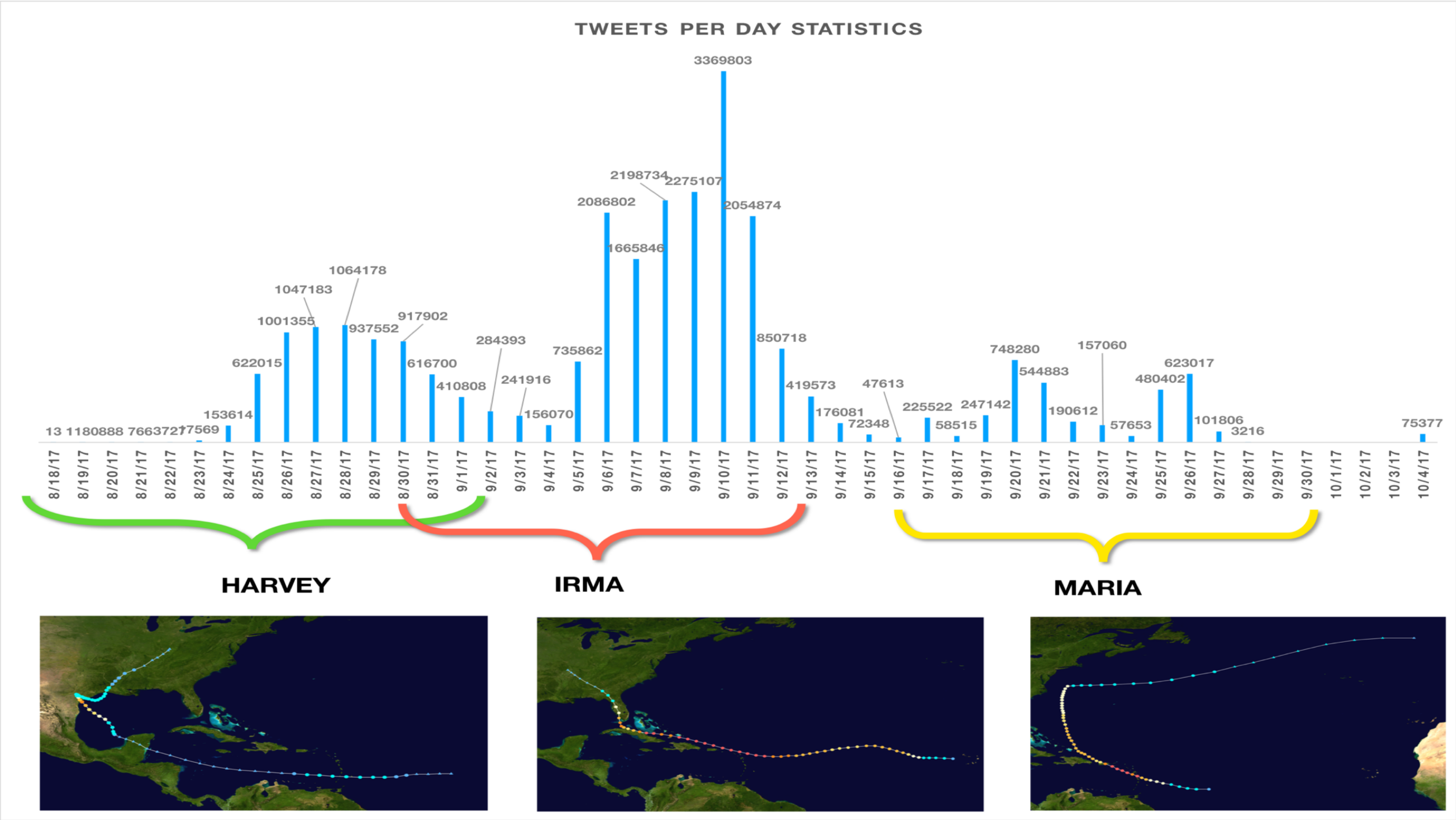
Domain adaptation [Li et al., 2015; Li et al., 2017, Alam et al., 2018, Mazloom et al., 2018]

- Knowledge from a prior source disaster is transferred to a target disaster

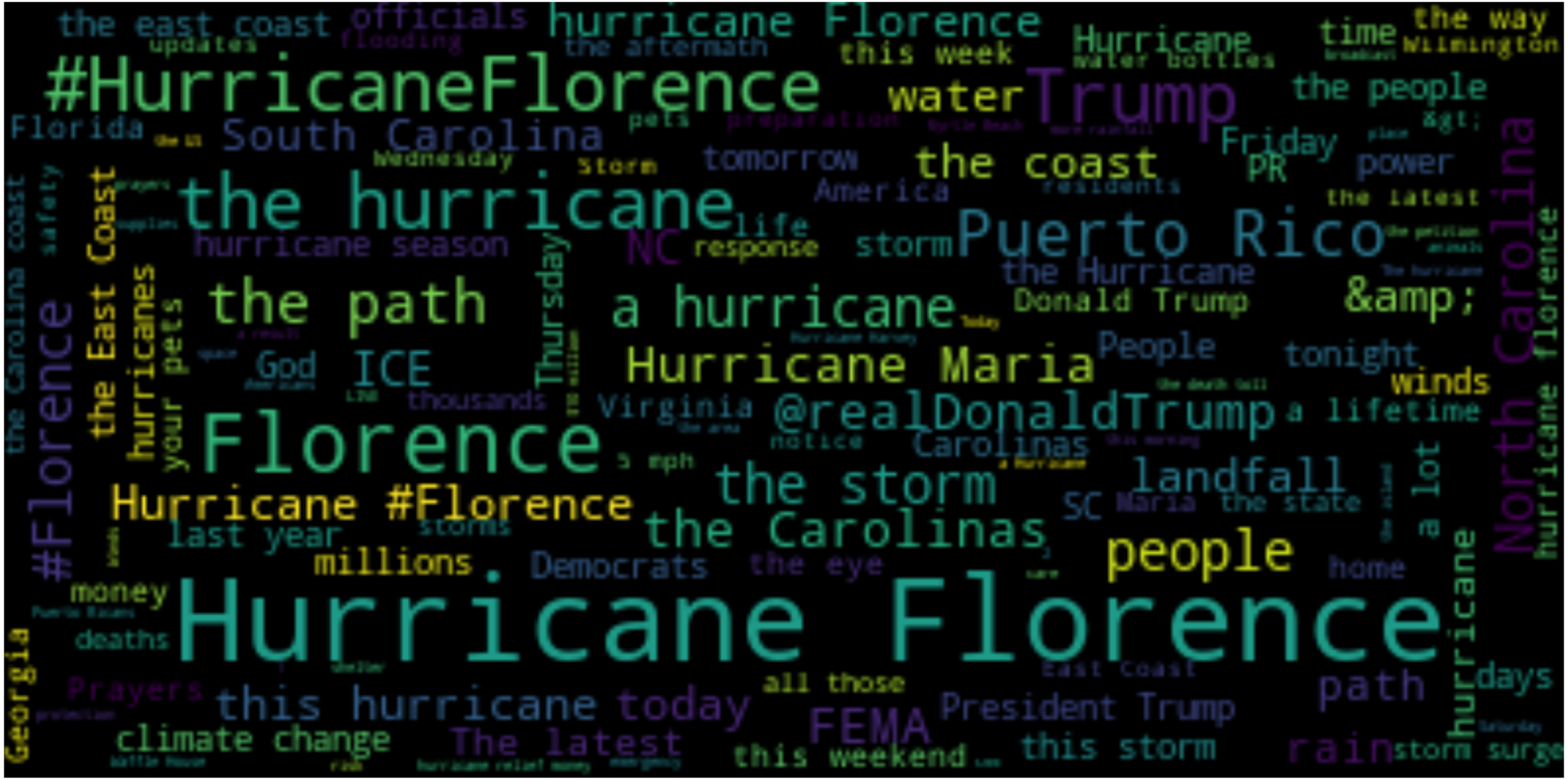
Unsupervised learning, e.g., topic modeling [Resch et al., 2017]

- Topic modeling can help associate topics/categories with tweets

Using Amazon Comprehend results to get aggregate statistics

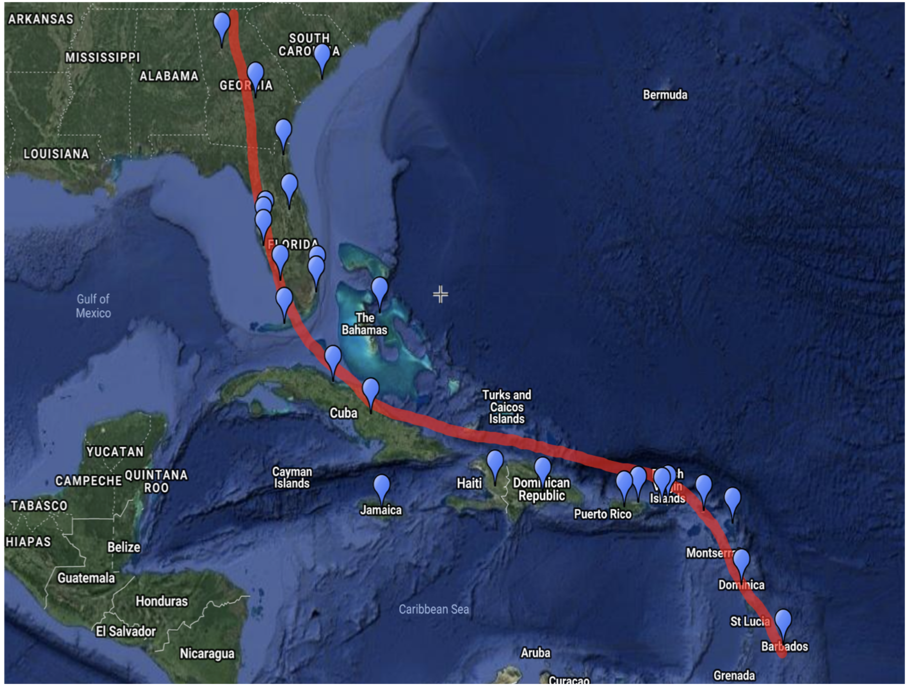


Using Comprehend results to determine frequent entities

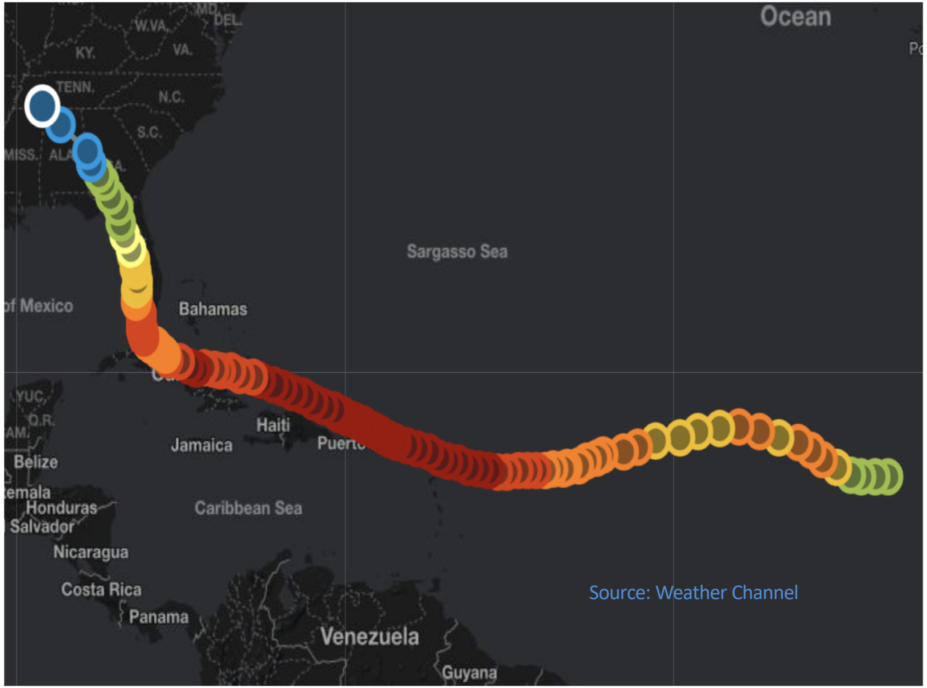


Using locations identified by Comprehend to track hurricane path

Hurricane Irma predicted path



Hurricane Irma real path



aws

Domain Adaptation Approaches for Classifying Crisis Related Data on Social Media

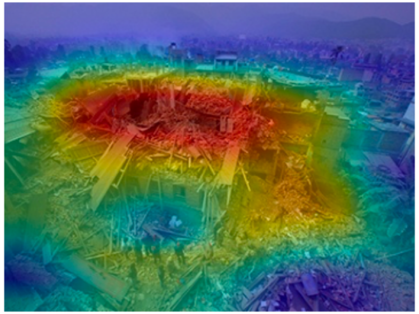
Doina Caragea, KSU

(1) Damage

(2) Damage

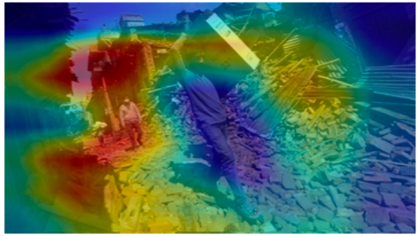
(3) Damage

(4) Damage

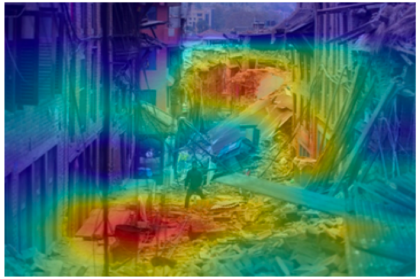


(1) DAV = 0.413

(2) DAV = 0.453



(3) DAV = 0.423



(4) DAV = 0.385

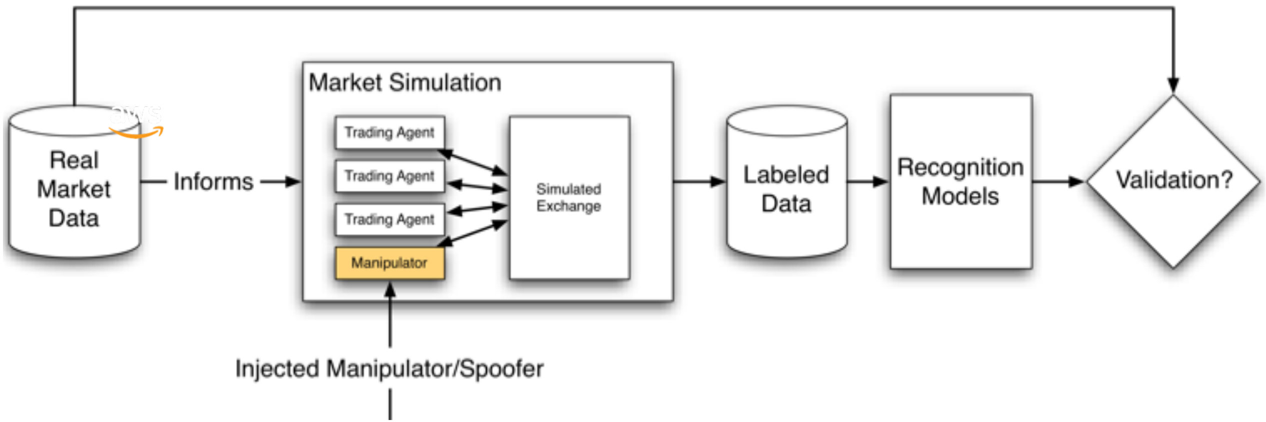
We thank the National Science Foundation and Amazon Web Services for support from the grant IIS-1741345.

IEEE/ACM ASONAM 2018, August 28-31, 2018, Barcelona, Spain
978-1-5386-6051-5/18/$31.00 © 2018 IEEE

BigData Market Manipulation Project

Michael Wellman, University of Michigan

- Collaboration between U.Michigan and Georgia Tech
- Sponsored by NSF BIGDATA program, computational support from AWS
- Interdisciplinary: Computer Science (AI/ML), Finance, Law & Public Policy



<u>Goal</u>: New techniques for detecting and mitigating manipulation

Market Manipulation

MARKETS

As 'Spoof' Trading Persists, Regulators Clamp Down

Bluffing Tactic That Dodd-Frank Banned in 2010 Can Distort Markets

WSJ's Bradley Hope explains how regulators are cracking down on "spoofing," a trading move designed to trick other investors into buying and selling at artificially high or low prices. Photo: Getty

By BRADLEY HOPE   70 COMMENTS

Updated Feb. 22, 2015 10:34 p.m. ET

CHICAGO—One June morning in 2012, a college dropout whom securities traders call "The Russian" logged on to his computer and began trading Brent-crude futures on a London exchange from his skyscraper office here.

Over six hours, Igor Oystacher's computer sent roughly 23,000 commands, including thousands of buy and sell orders, according to correspondence from the exchange to his clearing firm reviewed by The Wall Street Journal. But he canceled many of those orders milliseconds after placing them, the documents show, in what the exchange alleges was part of a trading practice designed to trick other investors into buying and selling at artificially high or low prices.

Traders call the illegal bluffing tactic "spoofing," and they say it has long been used to manipulate prices of anything from stocks to bonds to futures.

US seals first prosecution against stock market trader for 'spoofing'

A jury convicts Michael Coscia on six charges of commodities fraud and six charges of spoofing, all of the charges he faced

9   0   4   13   Email

Prosecutors said Michael Coscia wanted to lure other traders to markets by creating an illusion of demand so that he could make money on smaller trades Photo: AP

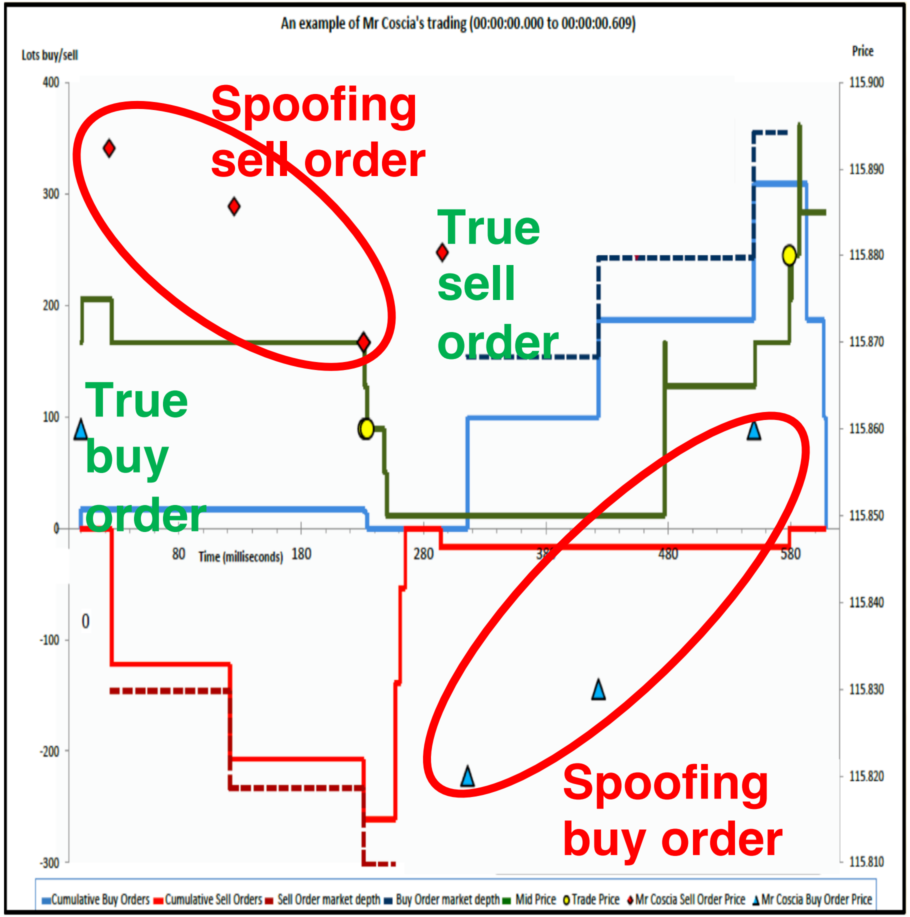
By Reuters

11:48PM GMT 03 Nov 2015

A US jury has found high-frequency trader Michael Coscia guilty of commodities fraud and "spoofing" in the US government's first criminal

Spoofing

Definition: Practice of submitting large spurious orders to buy or sell some security to mislead other traders about market state



Source: UK Financial Conduct Authority

Amazon SageMaker

Build

Pre-built notebook instances

Highly-optimized machine learning algorithms

Train

One-click training for ML, DL, and custom algorithms

Easier training with hyperparameter optimization

Deploy

Deployment without engineering effort

Fully-managed hosting at scale

Collaboration with NSF and Internet2: E-CAS

Internet2 and @NSF_CISE Announce Selection of First-Phase Research Proposals for Exploring Clouds for Acceleration of Science (E-CAS) Project https://bit.ly/2U56d1R #NSFfunded #research #cloud

Tweet this

The successful proposals for the year-long first phase of the E-CAS project are:

Development of BioCompute Objects for Integration into Galaxy in a Cloud Computing Environment
Raja Mazumder, George Washington University

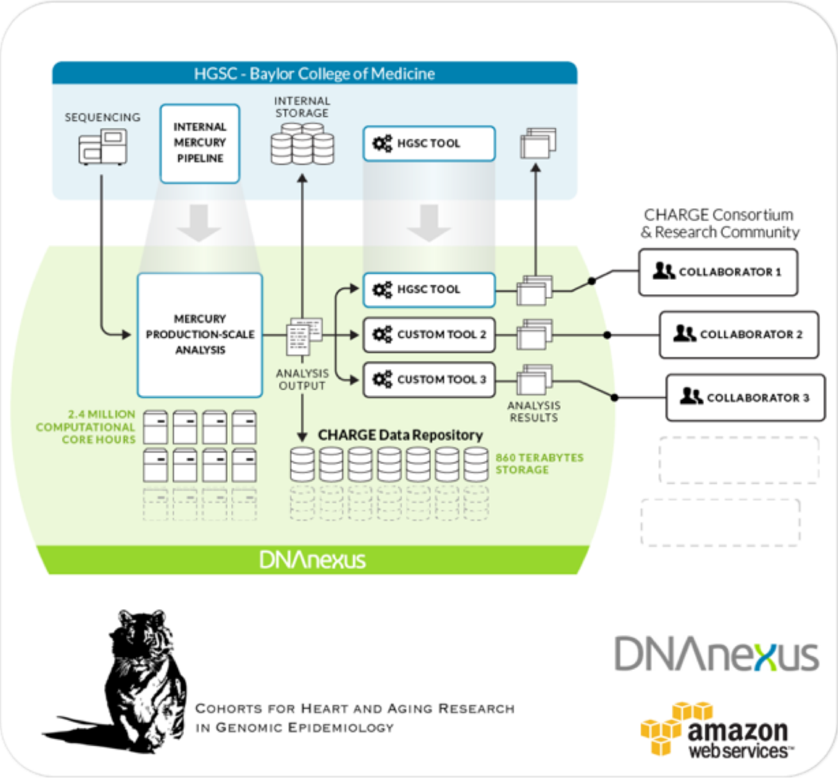
BioCompute objects allow researchers to describe bioinformatic analyses comprised of any number of algorithmic steps and variables to make computational experimental results clearly understandable and easier to repeat. This project will create a library of BioCompute objects that describe bioinformatic workflows on AWS, which can be accessed and contributed to by users of the widely used bioinformatics platform, Galaxy.

IceCube computing in the cloud
Benedikt Riedel, University of Wisconsin

The IceCube Neutrino Observatory located at the South Pole supports science from a number of disciplines including astrophysics, particle physics, and geographical sciences operating continuously being simultaneously sensitive to the whole sky. This project aims to burst into cloud to support follow-up computations of observed events, as well as alerts to and from the research community, such as other telescopes and LIGO.

https://www.businesswire.com/news/home/20190326005155/en/Internet2-National-Science-Foundation-Announce-Selection-First-Phase

Access to shared data repositories



Baylor College of Medicine CHARGE

Baylor College of Medicine Human Genome Sequencing Center and DNANexus using the Mercury Pipeline for the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium

Supports 300+ researchers around the world

Analyzed the genomes of over 14,000 individuals, encompassing 3,751 whole genomes and 10,940 whole exomes (~1PB of data)

Used 3.3 million core hours over 4 weeks to complete the job 5.7x faster than what could have been accomplished on-premise

The outcomes

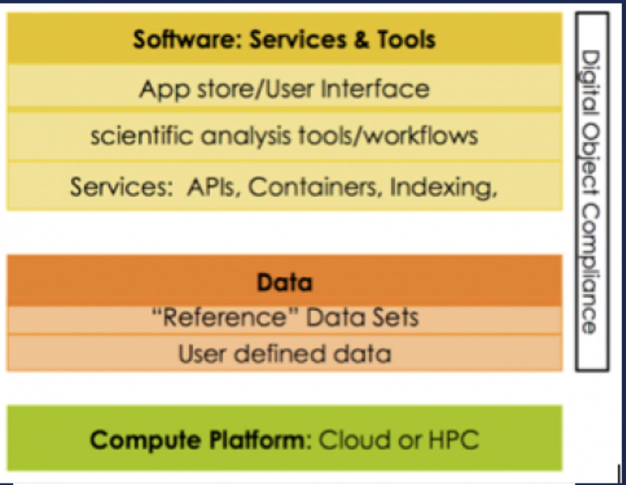
- Easier collaboration
- Faster time to science
- Cost-effective: On-premise was prohibitively expensive
- No longer constrained by on-premise capacity
- Scientists focusing on Science as opposed to infrastructure

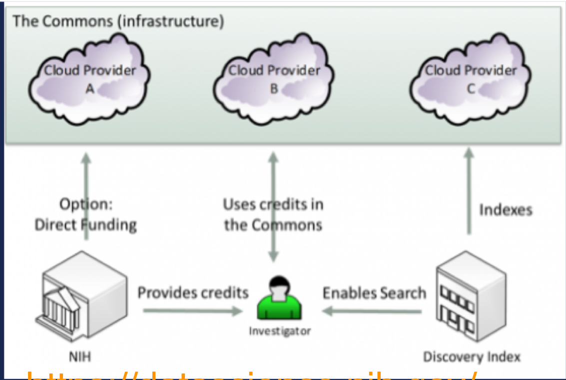
https://aws.amazon.com/solutions/case-studies/baylor/

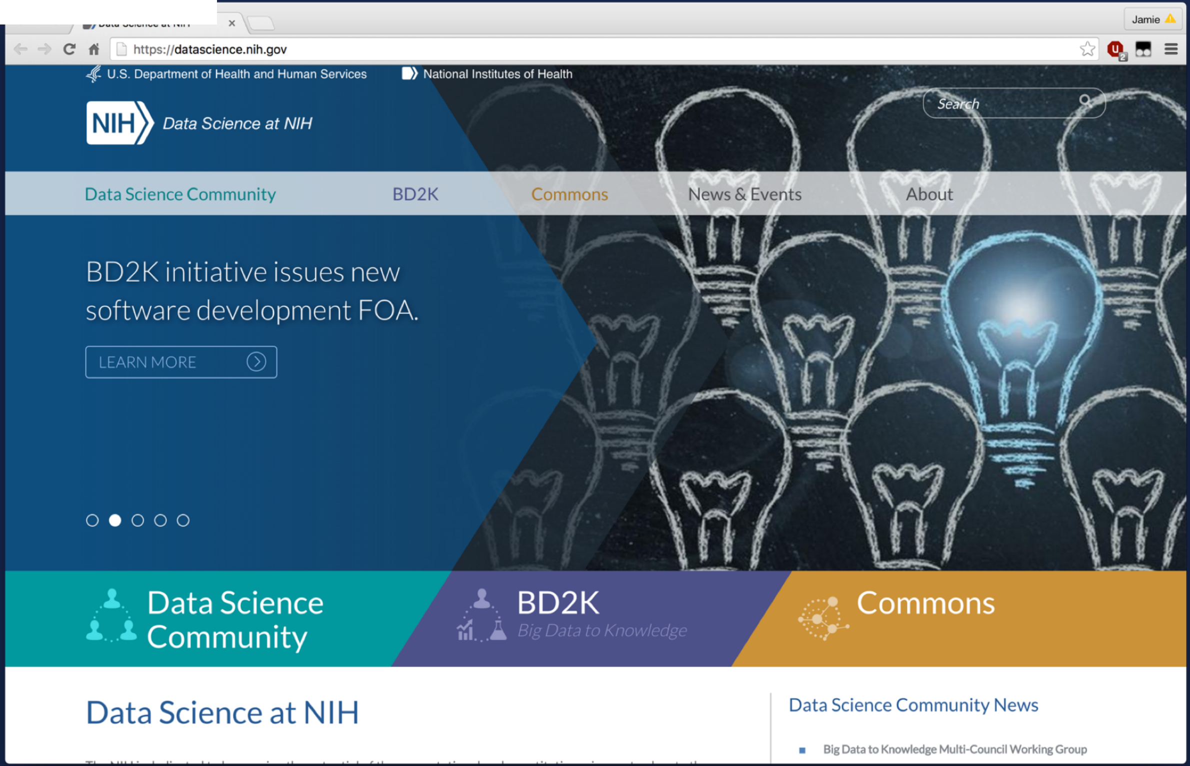
Harvard University: Precision Medicine using AWS

https://www.slideshare.net/AmazonWebServices/precision-medicine-on-the-cloud

National Institutes of Health







https://datascience.nih.gov/

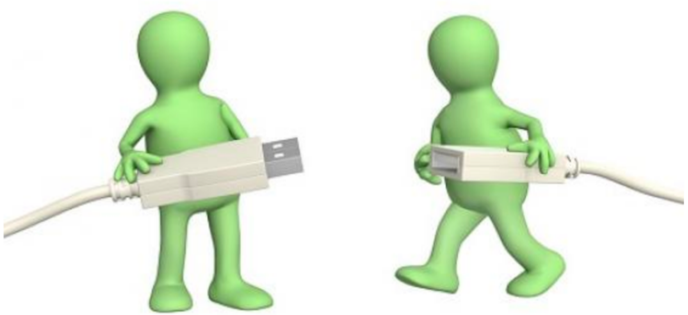
Supported by aws

aws SUMMIT

aws research initiative

Amazon Web Services joins NIH's STRIDES Initiative

October 24, 2018 | Danielle Brown | Analytics & Quality



AWS signs on with NIH cloud precision medicine project

Researchers associated with the National Institutes of Health (NIH) will have more resources to access and analyze data thanks to a new partnership with Amazon Web Services (AWS).

The newly-formed partnership with NIH's Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES) Initiative will give NIH-associated biomedical researchers access to AWS technologies. AWS is Amazon's cloud-based computing platform subsidiary.

Amazon And NIH To Link Biomedical Data And Researchers

Robin Seaton Jefferson Contributor

Retirement

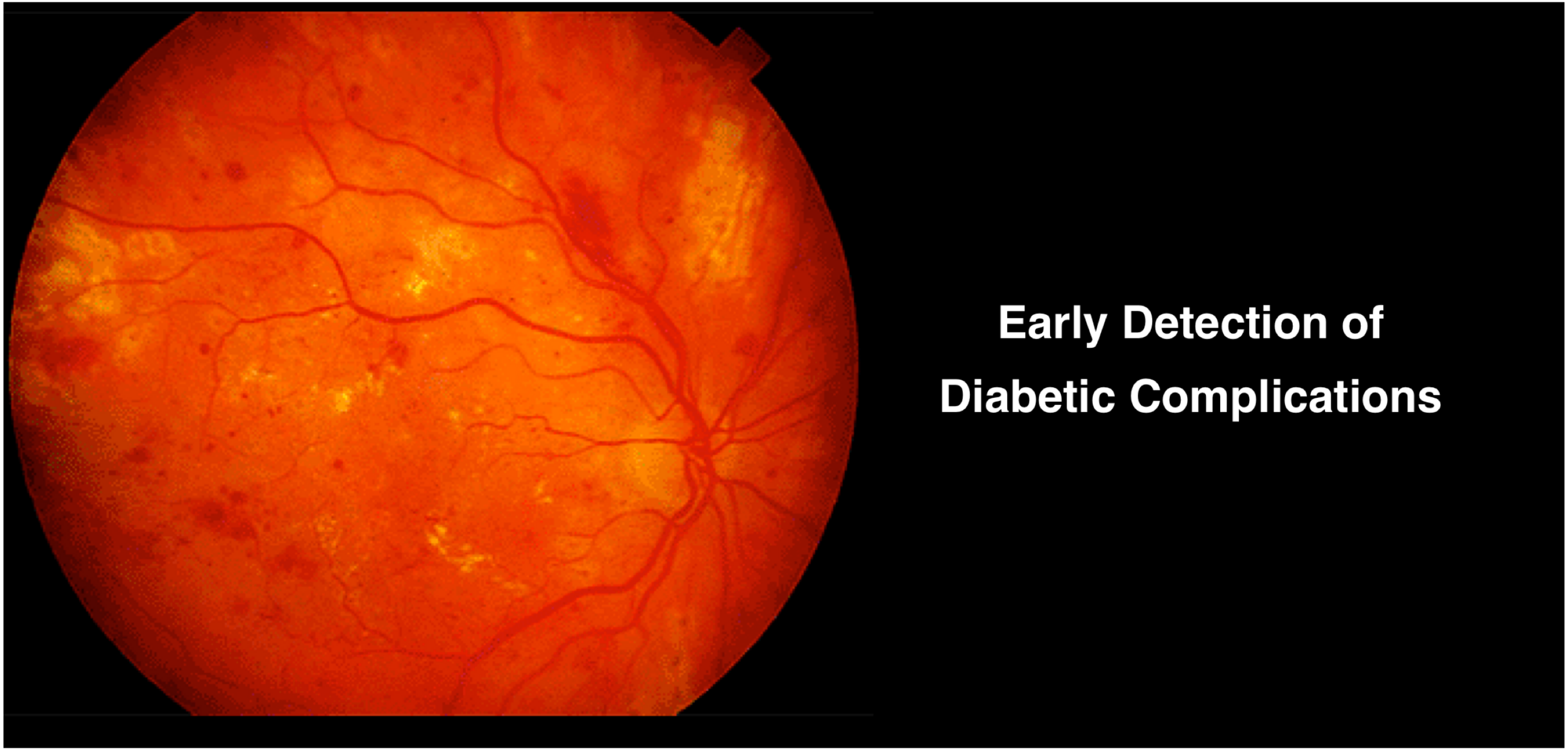
Getty

The National Institutes of Health (NIH) today announced the addition of Amazon Web Services (AWS) to its Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES) Initiative. Launched in July of this

https://aws.amazon.com/blogs/publicsector/aws-and-national-institutes-of-health-collaborate-to-accelerate-discoveries-with-strides-initiative/

https://www.forbes.com/sites/robinseatonjefferson/2018/10/23/amazon-and-nih-to-link-biomedical-data-and-researchers/

Stanford: Automatic Grading of Diabetic Retinopathy through Deep Learning using AWS

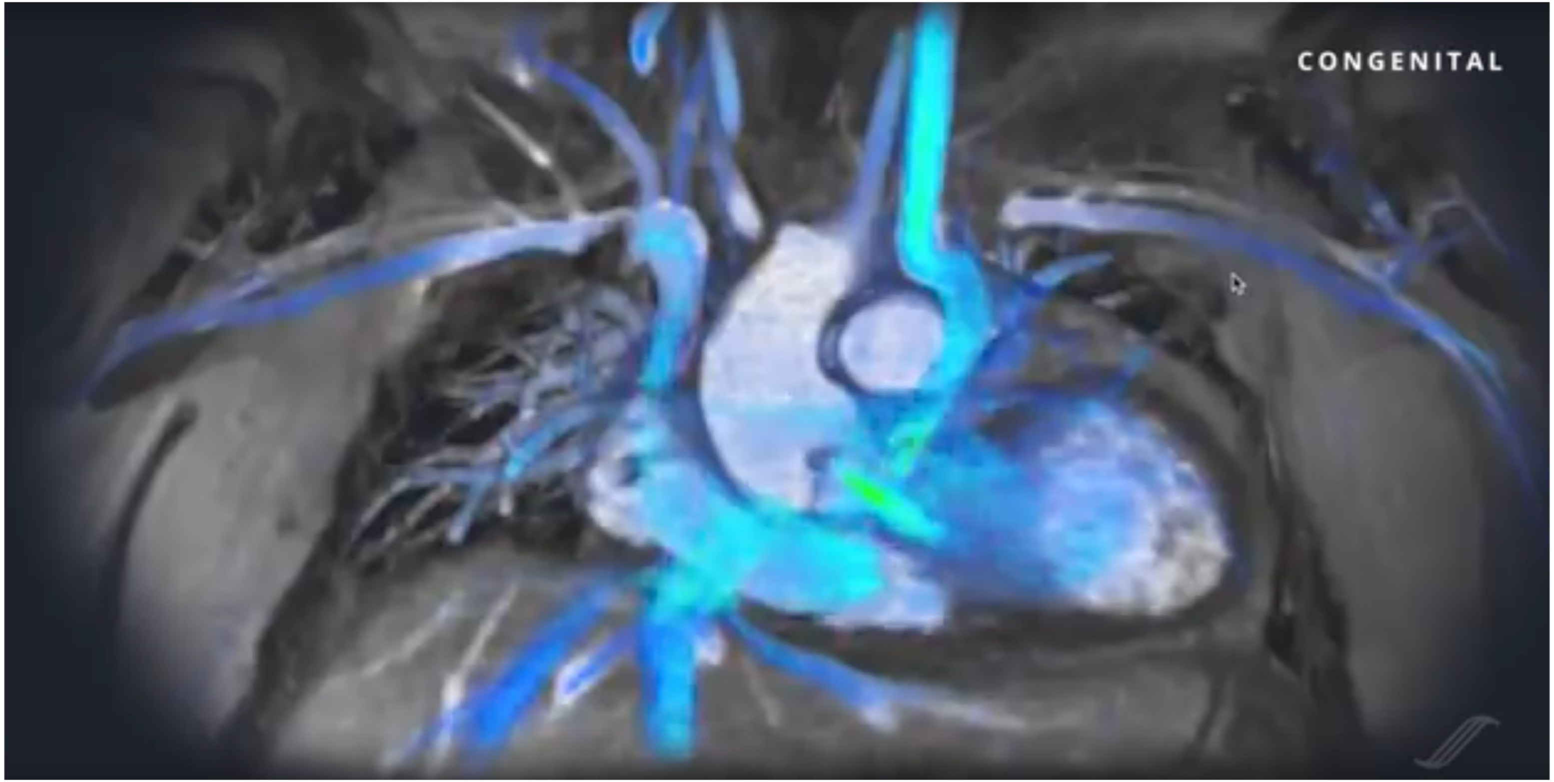


https://aws.amazon.com/blogs/publicsector/an-eye-on-science-how-stanford-students-turned-classwork-into-their-lifes-work/

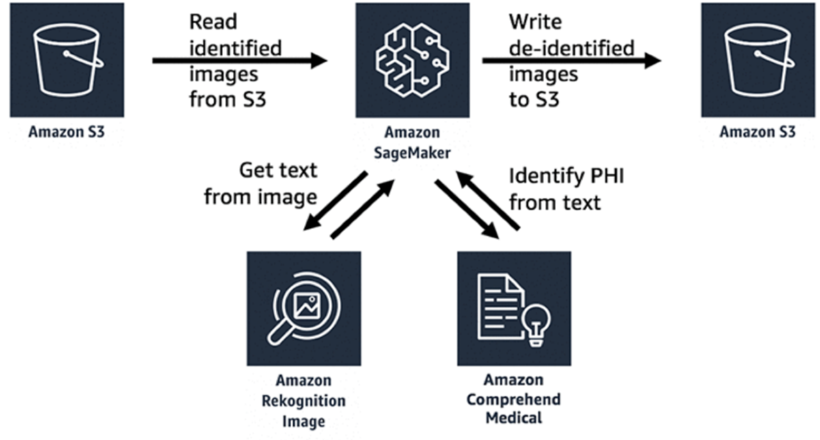
Skin Cancer Detection

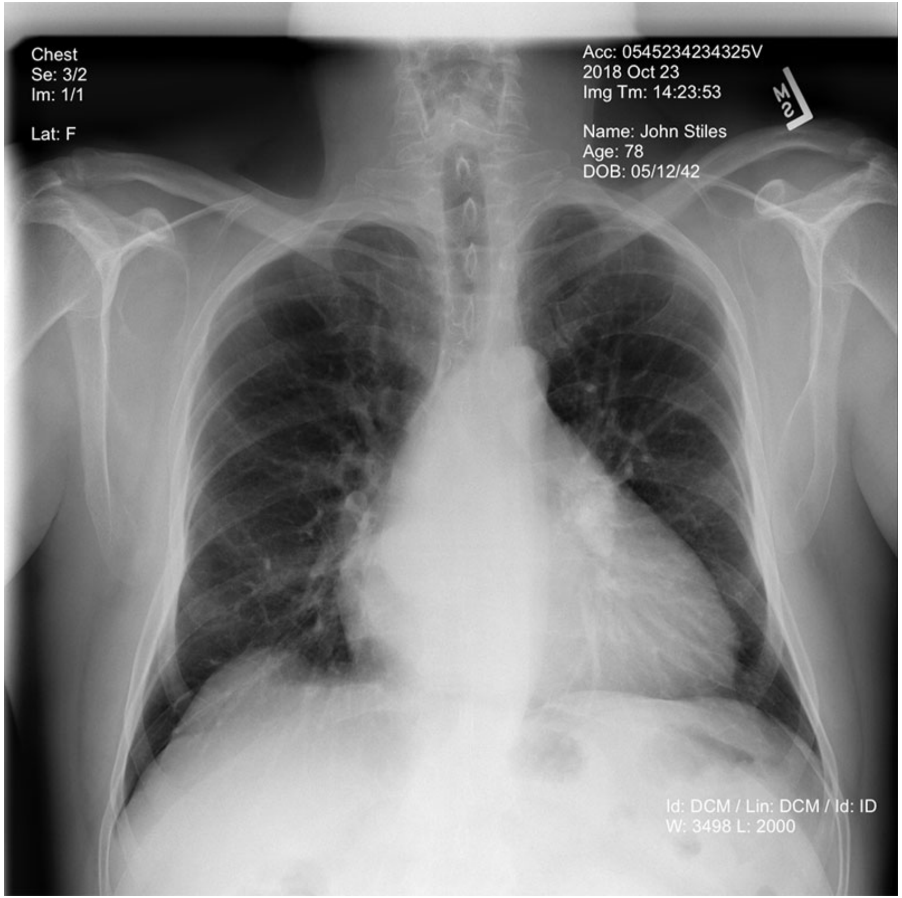
At Physician-Levels (or better)

FDA-Approved Medical Imaging

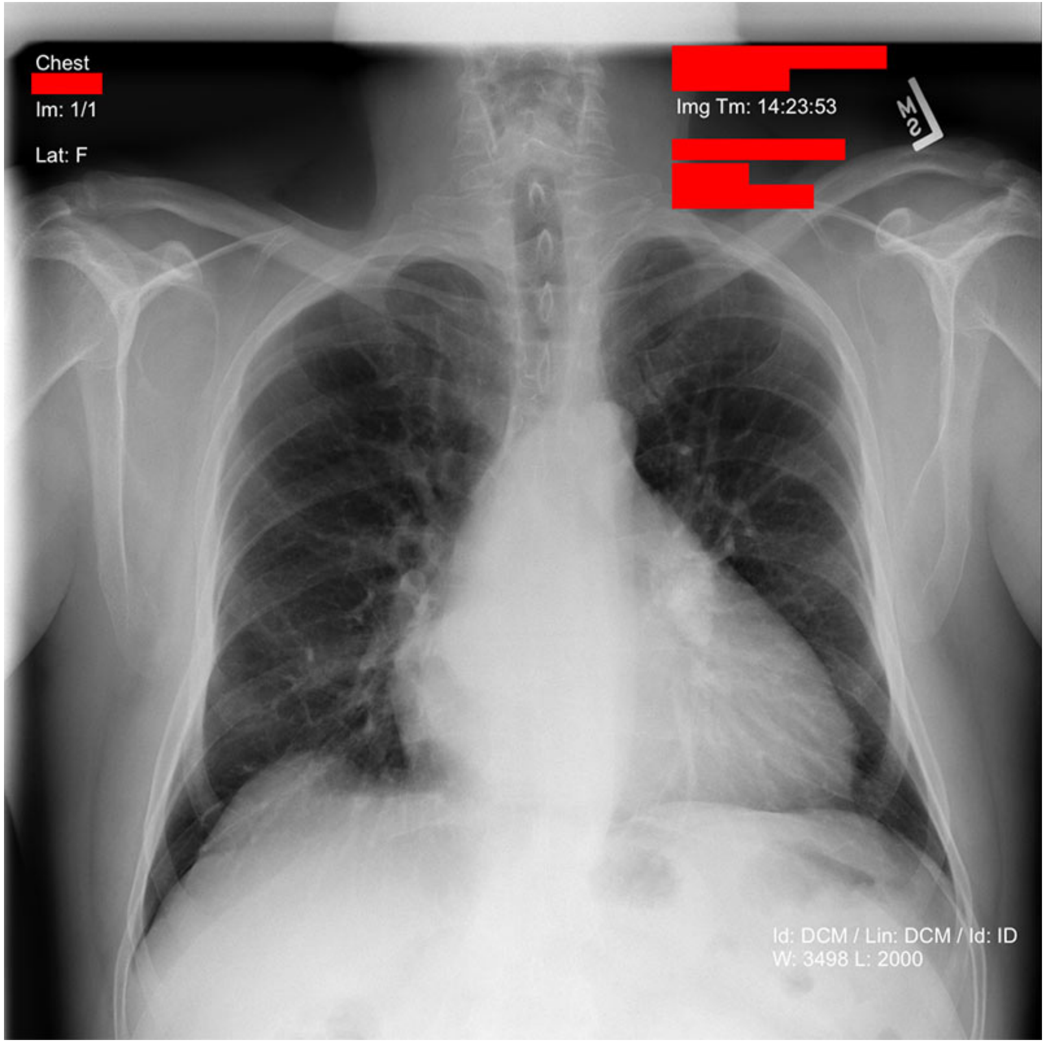


De-identify medical images with the help of Amazon Comprehend Medical and Amazon Rekognition





https://aws.amazon.com/blogs/machine-learning/de-identify-medical-images-with-the-help-of-amazon-comprehend-medical-and-amazon-rekognition/


Chest
Im: 1/1
Lat: F
Img Tm: 14:23:53
Id: DCM / Lin: DCM / Id: ID
W: 3498 L: 2000

amazon alexa

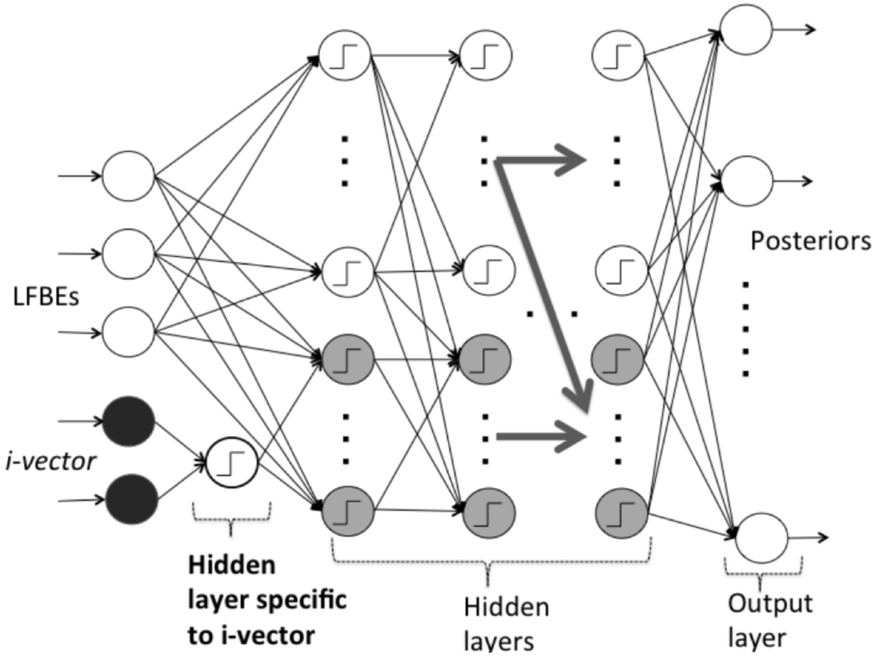
Language Support:

- English (Australia)
- English (Canada)
- English (India)
- English (UK)
- English (US)
- German
- Japanese

Speaker Adaptation

Speech Features

Speaker Features

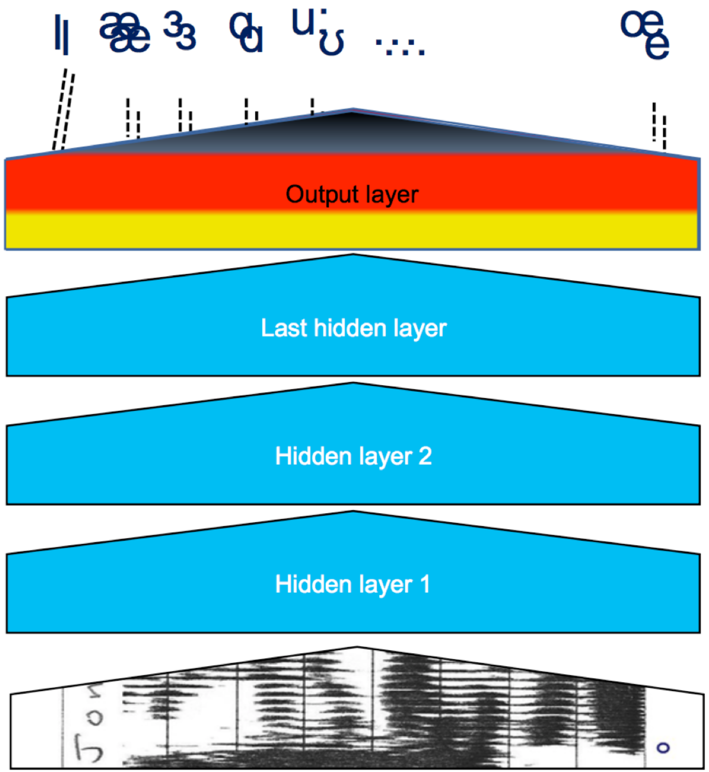


Phonetic Probabilities

5-7% relative reduction in word error rate compared to speaker independent model

Garimella, et al. "Robust i-Vector Based Adaptation of DNN Acoustic Model for Speech Recognition," Interspeech 2015

Transfer Learning from English to German



amazon alexa

Ohio Health Automates Patient Interactions

"Amazon Lex represents a great opportunity for us to deliver a better experience to our patients. Everything we do at OhioHealth is ultimately about providing the right care to our patients at the right time and in the right place. Amazon Lex's next generation technology and the innovative applications we are developing using it will help provide an improved customer experience. We are just scratching the surface of what is possible,"

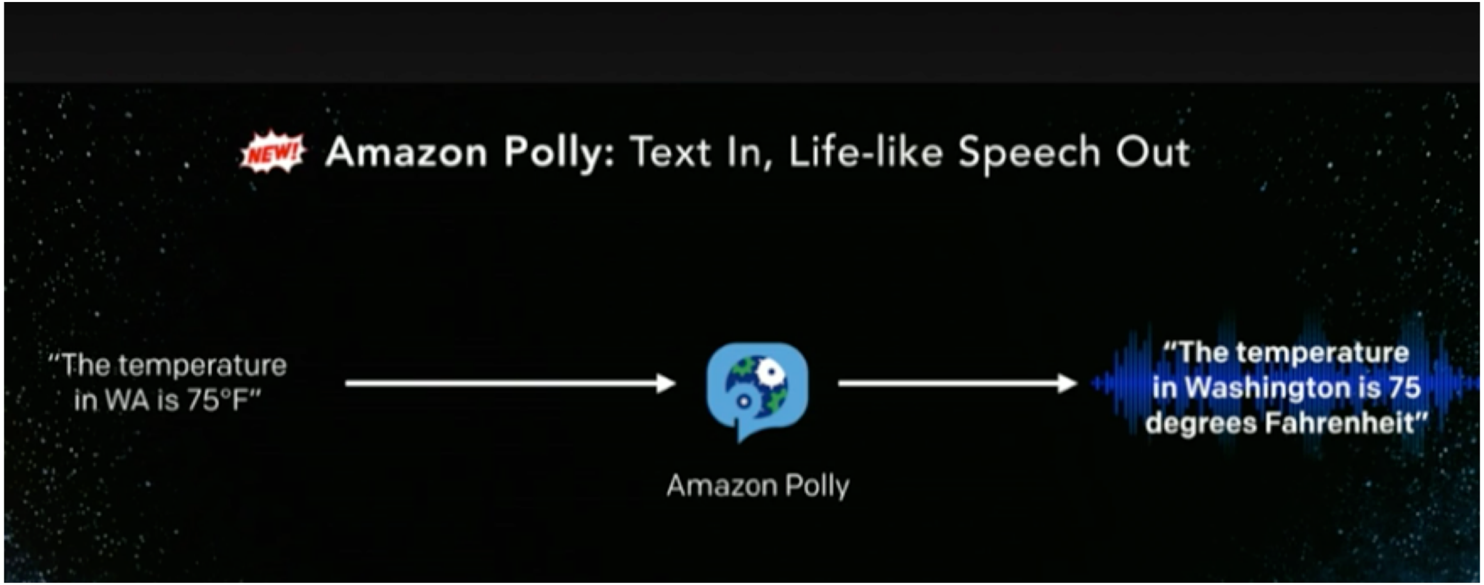
Michael Krouse
Senior Vice President and CIO – Ohio Health

OhioHealth
BELIEVE IN WE

- Delivers Personalized Care Recommendations
- Makes Customer Appointments
- Drives Urgent Care Referrals

Amazon Polly

Turn text into lifelike speech using deep learning



Incorporates ~47 different voices and fully managed services

https://aws.amazon.com/polly/

Amazon ML stack

AI SERVICES

(App developers with little knowledge of ML)

- **Vision**: Amazon Rekognition Image; Amazon Rekognition Video; Amazon Textract
- **Speech**: Amazon Polly; Amazon Transcribe
- **Language**: Amazon Translate; Amazon Comprehend
- **Chatbots**: Amazon Lex
- **Forecasting**: Amazon Forecast
- **Recommendations**: Amazon Personalize

ML SERVICES

(ML developers and data scientists)

Amazon SageMaker

- Ground Truth
- Notebooks
- Algorithms
- Marketplace
- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning
- Training
- Optimization (Neo)
- Deployment
- Hosting

ML FRAMEWORKS & INFRASTRUCTURE

(ML researchers and academics)

- **Frameworks**: TensorFlow; mxnet; PYTORCH; Chainer; HOROVOD
- **Interfaces**: GLUON; Keras
- **Infrastructure**: Amazon EC2 P3 & P3DN; Amazon EC2 C5; FPGAs; AWS Greengrass; Amazon Elastic Inference; Amazon Inferentia

aws

Deep Learning

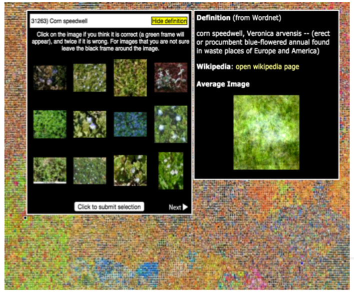
Significantly improve many applications on multiple domains

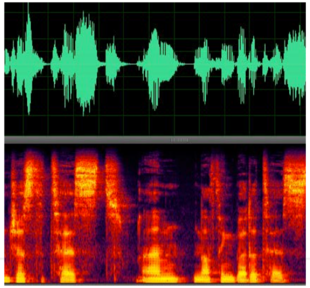
image understanding

speech recognition

natural language processing

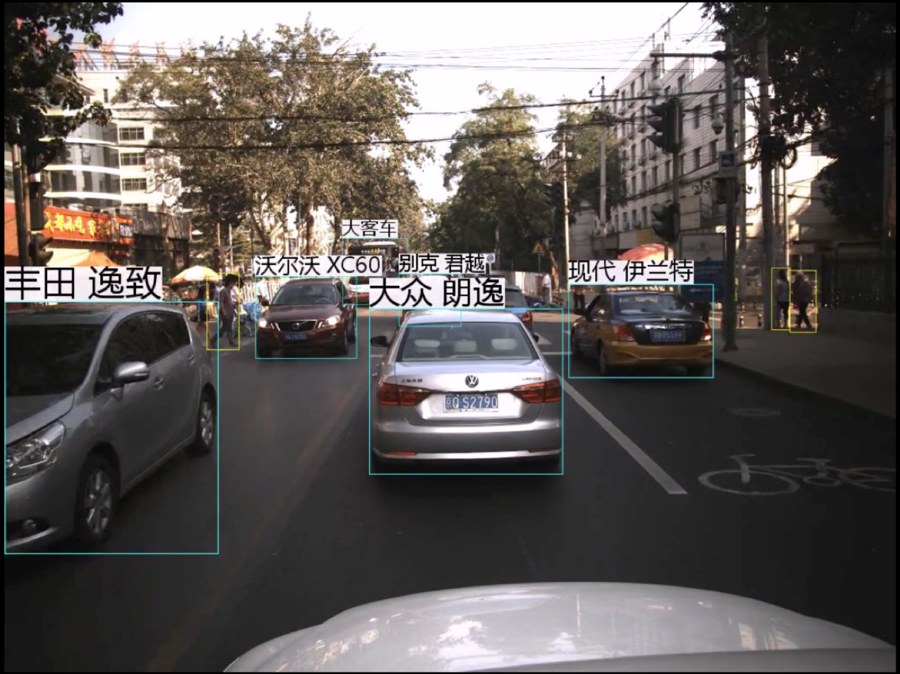
autonomy



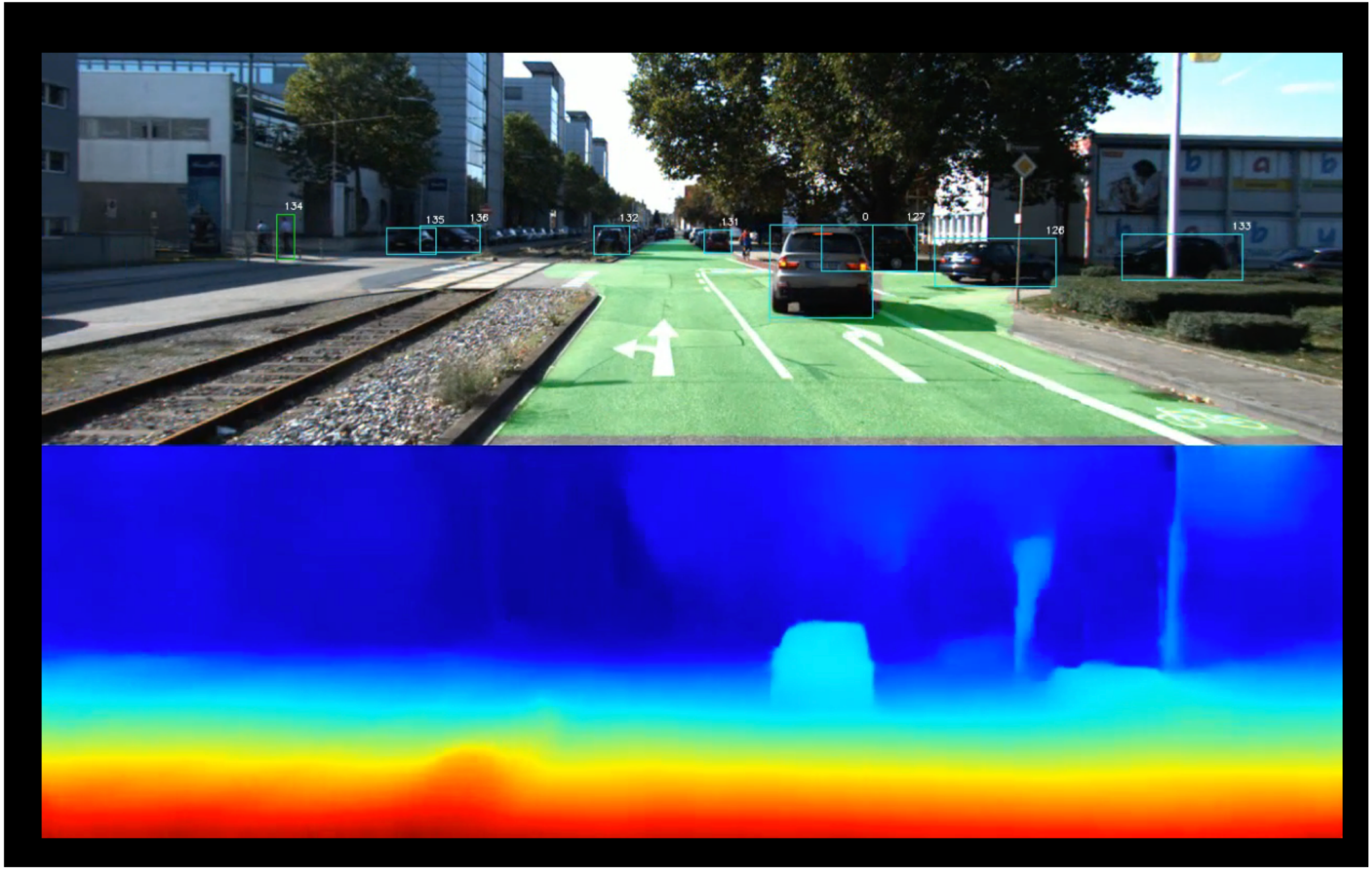


...

Autonomous Driving Systems



Centimeter-accurate positioning



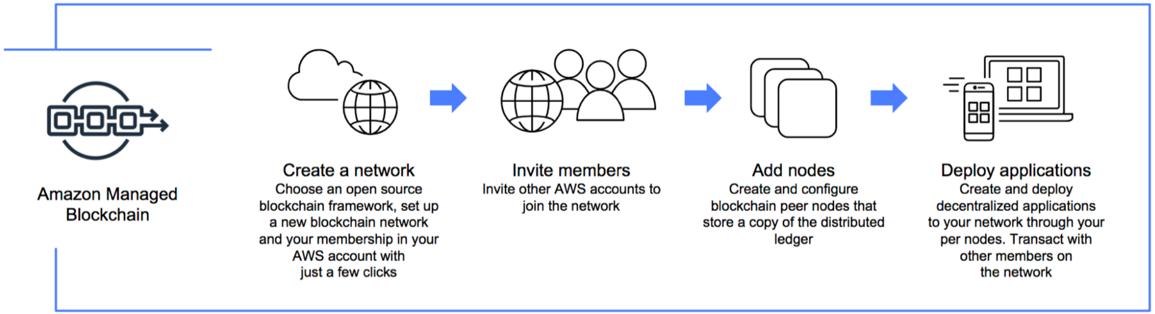
AWS DeepRacer

A fully autonomous 1/18th-scale race car designed to help you learn about reinforcement learning through autonomous driving

- Build machine learning models in Amazon SageMaker
- Train, test, and iterate on the track using the AWS DeepRacer 3D racing simulator
- Compete in the world's first global autonomous racing league, to race for prizes and a chance to advance to win the coveted AWS DeepRacer Cup

What is Amazon Managed Blockchain?

Amazon Managed Blockchain is a fully managed service that makes it easy to create and manage scalable blockchain networks using popular open source frameworks: **Hyperledger Fabric and Ethereum**



Applications in various domains

Proof of Ownership

Documents/Contracts

Digital Security Trading

Enterprise Platforms

Mortgage Loans

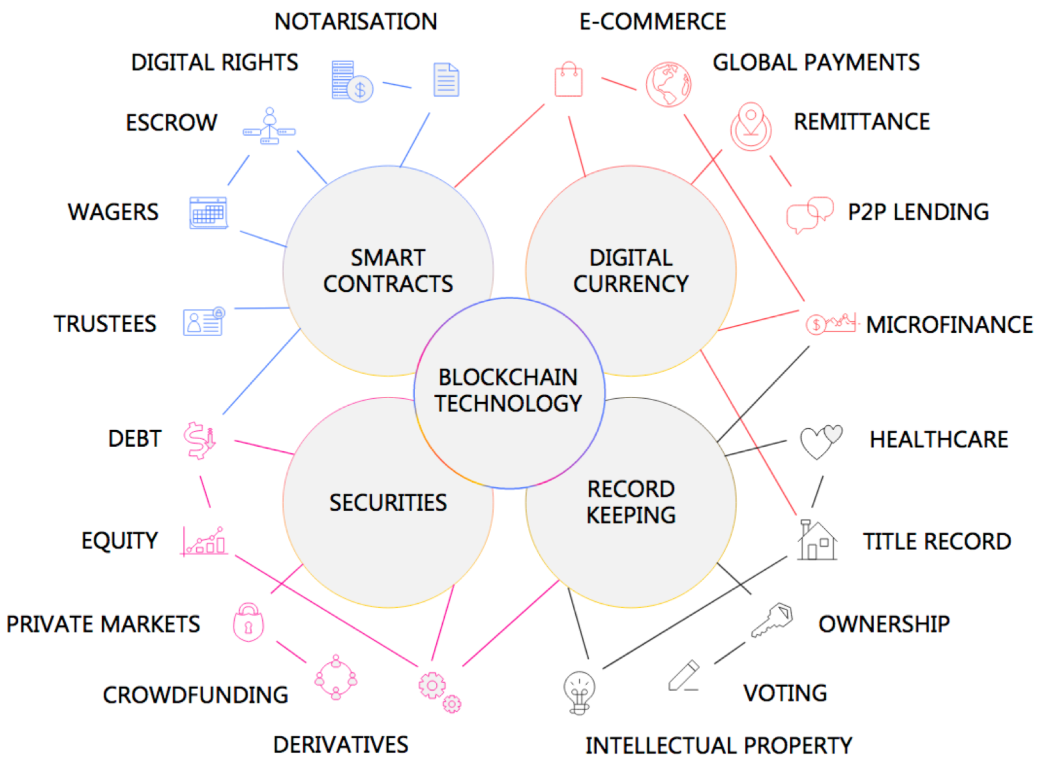
Voting Mechanisms

Patient Records

Corporate Governance

Financial

Insurance



Capital Markets

HCLS

Real Estate

Legal

Agriculture

Gaming

Transportation

M & E

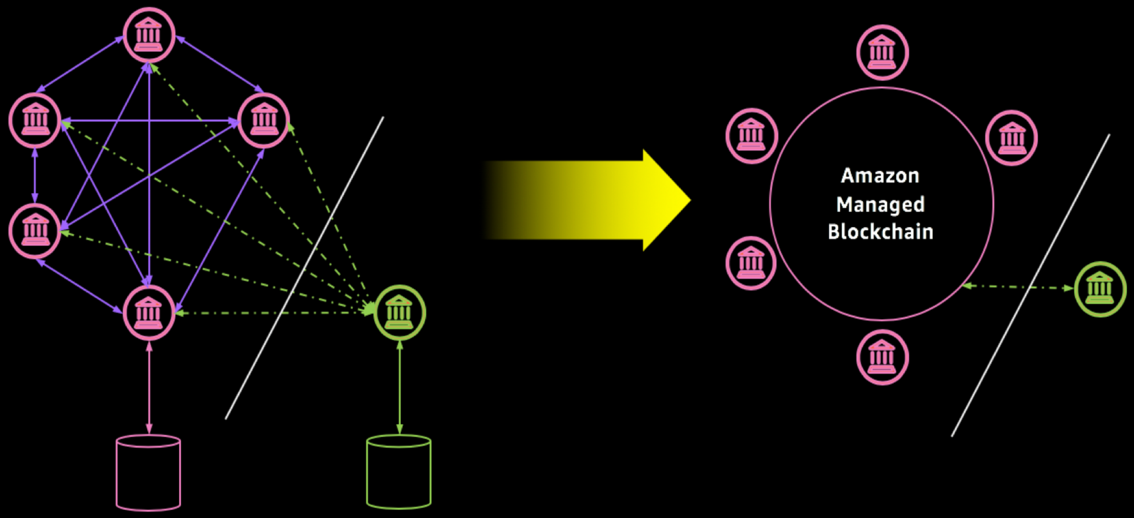
Digital Advertising

Power/Utilities

Retail

Cloud

Singapore Exchange: Project Ubin's blockchain use case



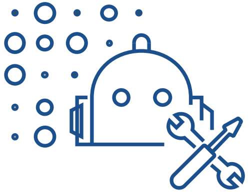
Challenges with existing financial systems:

- Lack of trust
- Inefficient processes for sending data across borders
- API divergence is expensive and cumbersome to maintain

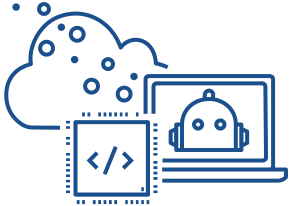
Benefits of implementing a blockchain

- Distributed application provides trust
- Provides reliability and resiliency
- Easy to add new participating members
- Efficient transfer of data and transactions without intermediaries

AWS RoboMaker - Robotics

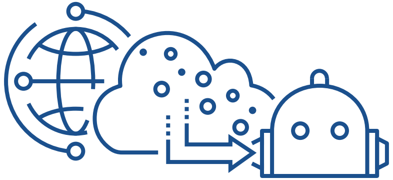


Cloud Extensions
for ROS



Development
Environment

Simulation



Fleet
Management

ASU Arizona State University

Cal Poly

Dartmouth

Northeastern University

RMIT University

Rutgers The State University of New Jersey

UCF University of Central Florida

University of Cambridge

The University of Texas at Austin Computer Science

W University of Washington

University of Waterloo

aws

AWS Cloud Credits for Research

1. Build cloud-hosted publicly available science-as-a-service applications, software, or tools to facilitate their future research and the research of their community.
2. Perform proof of concept or benchmark tests evaluating the efficacy of moving research workloads or open data sets to the cloud.
3. Train a broader community on the usage of cloud for research workloads via workshops or tutorials.

AWS Machine Learning Research Awards

Funding academic research at the forefront of machine learning.

amazon research awards

aws educate

https://aws.amazon.com/research-credits/

https://aws.amazon.com/aws-ml-research-awards/

https://ara.amazon-ml.com/proposals/#apply

https://aws.amazon.com/opendata/

Open Data on AWS

Share any volume of data with as many people as you want

aws

AWS Research Workshops

This repo provides a managed SageMaker jupyter notebook with a number of notebooks for hands on workshops in data lakes, AI/ML, Batch, IoT, and Genomics.

Workshops

> Please review and complete all prerequisites before attempting these workshops.

Title	Description
Introduction to AWS Basics	Learn about core AWS services for compute, storage, database and networking. This workshop has a hands-on lab where you will be able to launch an auto-scaled Apache web server behind an ALB, S3 bucket hosting content of the home page, and how to define the approriate roles for each resource.
Buidling a Data Lakes	In this series of hands-on workshops, you will learn how to understand what data you have, how to drive insights, and how to make predictions using purpose-built AWS services. Learn about the common pitfalls of building data lakes, and discover how to successfully drive analytics and insights from your data. Also learn how services such as Amazon S3, AWS Glue, Amazon Athena, and Amazon AI/ML services work together to build a serverless data lake for various roles, including data scientists and business users.
Build Serverless Applications in Python with AWS SAM CLI **Coming Soon**	AWS Serverless Applications in Python: With AWS Serverless computing you can run applications and services without having to provision, scale, and manage any servers. In this workshop, we will introduce the basics of building serverless applications and microservices using services like AWS Lambda, Amazon API Gateway, Amazon DynamoDB, and Amazon S3. You'll learn to build and deploy your own serverless application using these services for common use cases like web applications, analytics, and more.
Tensorflow with Amazon SageMaker	Amazon SageMaker is a fully- managed platform that enables developers and data scientists to quickly and easily build, train, and deploy machine learning models at any scale. Amazon SageMaker removes all the barriers that typically slow down developers who want to use machine learning. We will show you how to train and build a ML model on SageMaker then how to deploy the inference end points on tools like AWS Greengrass or Serverless applications.

https://github.com/aws-samples/aws-research-workshops

Amazon SageMaker Examples

This repository contains example notebooks that show how to apply machine learning and deep learning in Amazon SageMaker

Examples

Introduction to Ground Truth Labeling Jobs

These examples provide quick walkthroughs to get you up and running with the labeling job workflow for Amazon SageMaker Ground Truth.

- From Unlabeled Data to a Deployed Machine Learning Model: A SageMaker Ground Truth Demonstration for Image Classification is an end-to-end example that starts with an unlabeled dataset, labels it using the Ground Truth API, analyzes the results, trains an image classification neural net using the annotated dataset, and finally uses the trained model to perform batch and online inference.
- Ground Truth Object Detection Tutorial is a similar end-to-end example but for an object detection task.
- Basic Data Analysis of an Image Classification Output Manifest presents charts to visualize the number of annotations for each class, differentiating between human annotations and automatic labels (if your job used auto-labeling). It also displays sample images in each class, and creates a pdf which concisely displays the full results.
- Training a Machine Learning Model Using an Output Manifest introduces the concept of an "augmented manifest" and demonstrates that the output file of a labeling job can be immediately used as the input file to train a SageMaker machine learning model.
- Annotation Consolidation demonstrates Amazon SageMaker Ground Truth annotation consolidation techniques for image classification for a completed labeling job.

Introduction to Applying Machine Learning

These examples provide a gentle introduction to machine learning concepts as they are applied in practical use cases across a variety of sectors.

https://github.com/awslabs/amazon-sagemaker-examples

Thank you!

Sanjay Padhi

sanpadhi@amazon.com