Pacific Northwest NATIONAL LABORATORY

Applications of Machine Learning to ARM/ASR Science

June 17, 2019

Conveners: Joseph Hardin, Rao Kotamarthi, Jennifer Comstock, Shaocheng Xie, Ed Luke

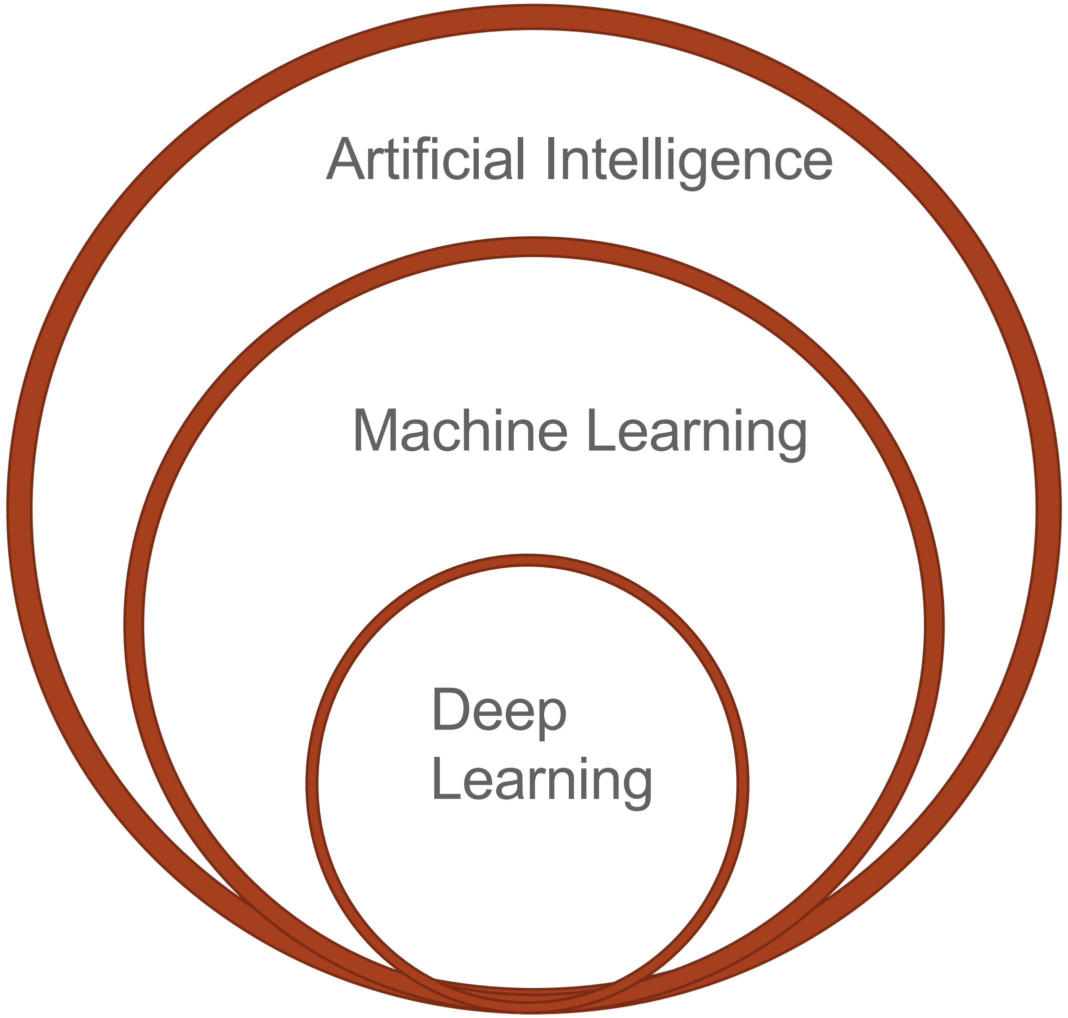
U.S. DEPARTMENT OF ENERGY BATTELLE

PNNL is operated by Battelle for the U.S. Department of Energy

Pacific Northwest NATIONAL LABORATORY

Machine Learning

- Machine learning is a subfield of artificial intelligence that focuses on designing systems that can “learn” from data to solve problems without requiring explicit programming of those solutions.
- Utilizes techniques from electrical engineering, computer science, mathematics, and statistics to solve a variety of problems in every field.
- Has great potential to revolutionize science by uncovering previously hidden correlations and predictors.
- Used in everything from climate models to self driving cars.



COLUMBIA NEWS

Video Books News Archive Press Room

Home > News Archive > Combating Climate Change With Artificial Intelligence

RESEARCH

Combating Climate Change With Artificial Intelligence

re-

scientist team up to apply modern

YAHOO! FINANCE

Search for news, symbols or companies

inance Home Watchlists My Portfolio Screeners Premium

&P 500	Dow 30	Nasdaq
878.46	26,010.77	7,749.97
34.97 (+1.23%)	+290.11 (+1.13%)	+134.42 (+1.77

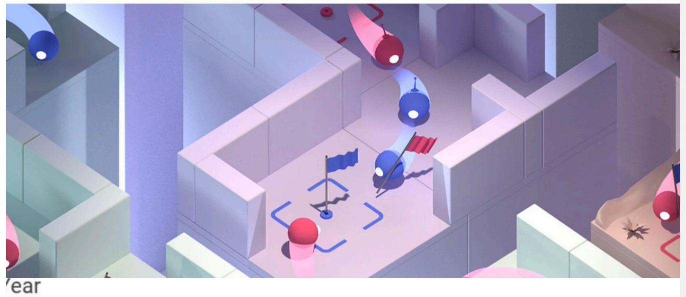
IBM Weather Signals Uses AI to Enable Predictive Weather-based Business Forecasting

CISION PR Newswire May 22, 2019

The New York Times

DeepMind Can Now Beat Us at Multiplayer Games, Too

Chess and Go were child's play. Now A.I. is winning at capture the flag. Will such skills translate to the real wo



"Arctic Siren" –AI Reveals Unexpected Source of Extreme Weather

Posted on May 29, 2019 in Climate Change, Featured Articles, Science, Technology

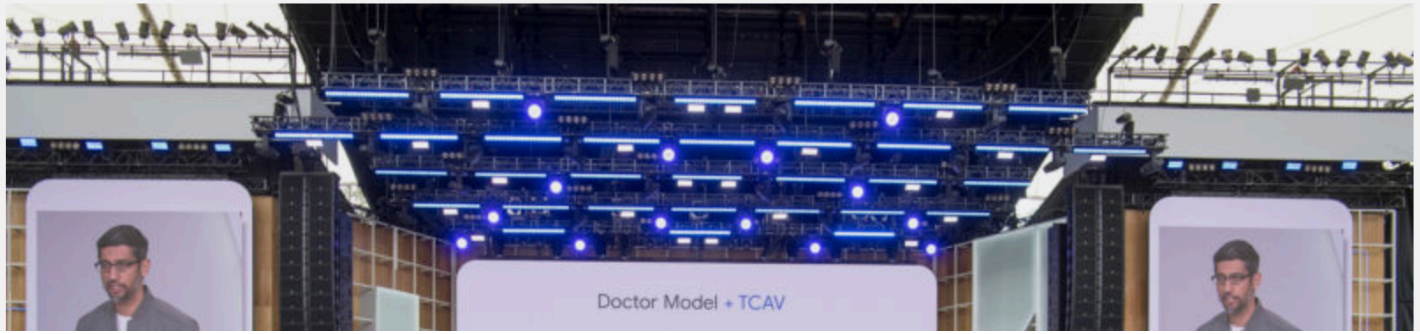
ars TECHNICA BIZ & IT TECH SCIENCE POLICY CARS GAMING & CULTURE STORE

AI FOR EVERYONE —

Why Google believes machine learning is its future

We heard so much about machine learning at Google I/O this year.

TIMOTHY B. LEE - 5/10/2019, 12:15 PM



Machine learning in ARM and ASR

- ARM and ASR have funded several machine learning projects.
  - We'll hear from some exciting projects today.
- Machine learning is already making an impact however from proxy models to data quality pushes.
- We felt it was important to get practitioners of machine learning in ARM/ASR together to discuss approaches and limitations and build a community.
- Discuss resources and opportunities that currently exist, and where we feel challenges are moving forward.

ARM/ASR Machine Learning Resources

- Data
  - ARM is one of the largest sources of atmospheric science data in the world.
  - Covers a wide variety of instruments, regimes, times, and campaigns.
- Stratus and Cumulus
  - Two institutional HPC clusters used by ARM and ASR.
  - Provides large
- Expertise
  - There are many people around ARM and ASR with expertise in machine learning
  - I encourage you to e-mail the presenters and conveners of this session and we can put you in contact with those that can help.

Stratus

- Stratus is a
  - 1080-core 30-node Cray cluster
  - 7.68 GB DDR4 memory **per core**
  - Each node consist of two 18 core processors.
  - 57.6 TB fast SSD
  - 100 TB Lustre storage.
  - InfiniBand network
  - Two systems with NVIDIA Kepler3 K80 (GK210) GB GPUs
- **Data adjacent**
  - Easily stage large quantities of ARM data to the system.
- Can be accessed as part of an ASR grant, or as part of ARM infrastructure
- Contact Giri or Jitu for more information on access and Stratus.

Types of machine learning overview

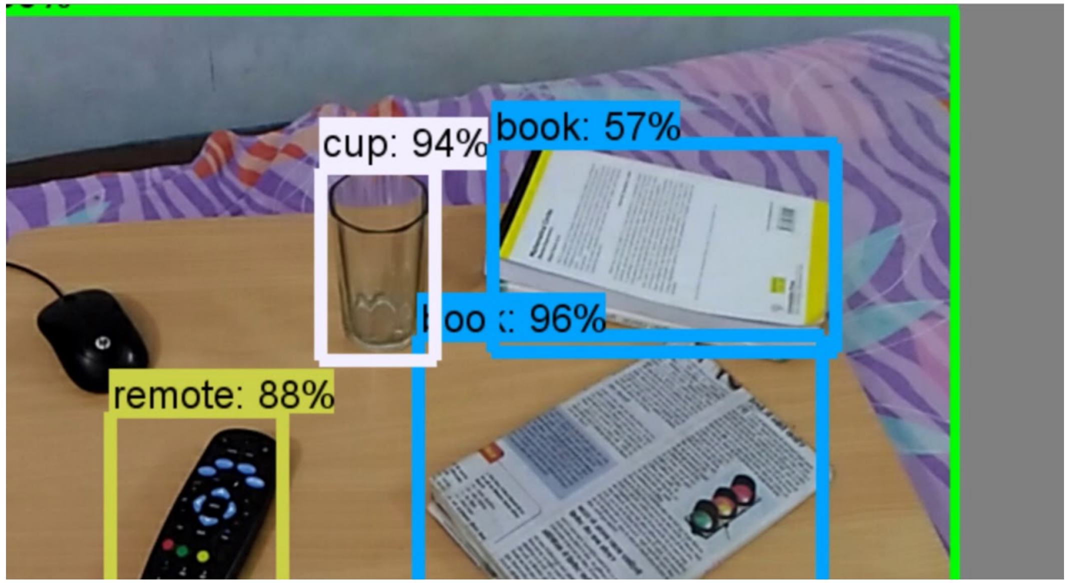
- Machine learning can be broadly classified into several different families.
  - If we think about ML as fitting a function Y=f(X) then the families are split by which parts we have (Y,X,f).

- This division has large implications for which types of problems can be tackled by each family

Supervised Learning

$$Y = f(X)$$

- Given a set of data points, and their corresponding output, find an f() capable of capturing the relation.
- Main difficulty is in labeling of data and finding representative features.
- Research focuses around better models and training methods.

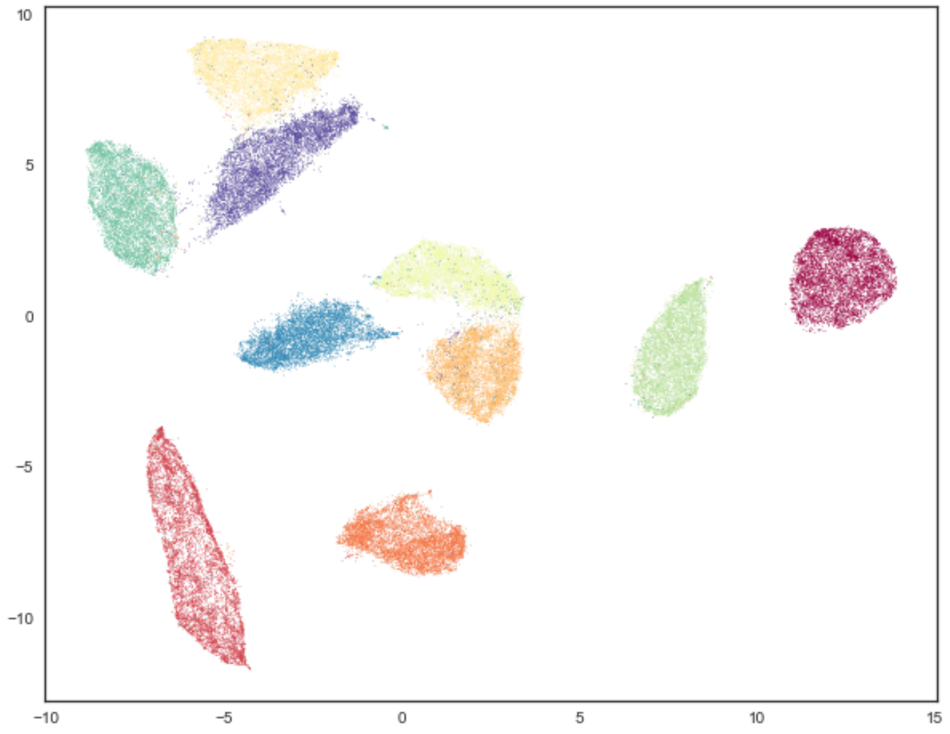


Source: youtube.com

Unsupervised Learning

X

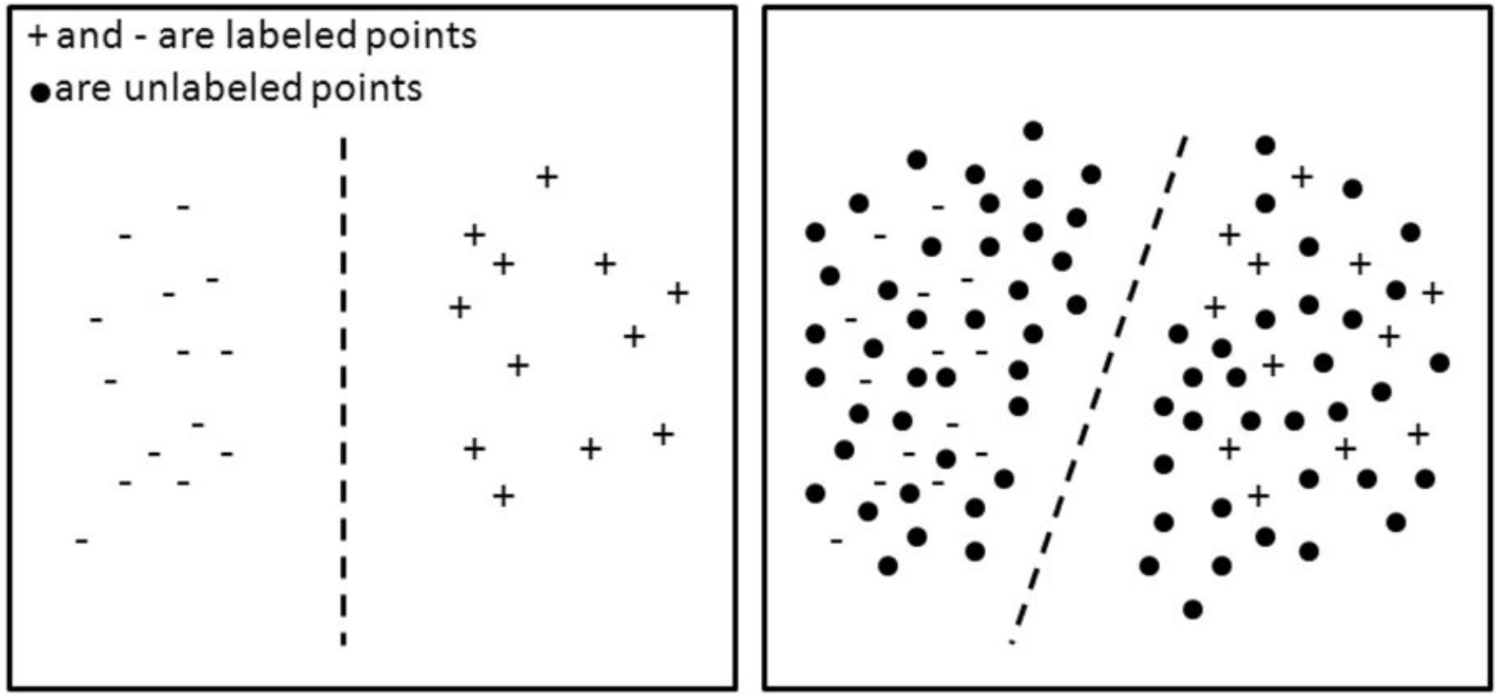
- When data is available, but without labels (Y), then training is “unsupervised”.
- Typical uses involve clustering, dimensionality reduction, and anomaly detection.
- Utilize statistics of data points to find a relation f that separate the data into:
  - Lower dimensional surface
  - Set of clusters



UMAP documentation

Semi-Supervised learning

- When you have a large set of data but only some samples are labeled training is “semi-supervised”
- Useful when labeling can be done, but the cost is large
- Can run training, and then get more label more data based on performance.
- Often used for data augmentation, directed attention, and other similar tasks



(a) (b)

Peikari et. al., ” A Cluster-then-label Semi-supervised Learning Approach for Pathology Image Classification“, Nature scientific Reports

Pacific Northwest NATIONAL LABORATORY

Reinforcement Learning

- Reinforcement-Learning
  - No a-priori data but you have a system that can make data
  - Used in game and simulation systems
- Behind recent outbreak of computers wining at various video games.
- Need a way of scoring performance of an algorithm for a task, and then you point it at a data source (simulator, video game, etc) and it iteratively improves.

Dexerto.com

Some applications of machine learning

- Proxy Models
- Parameterizations
- Data Fusion
- Super Resolution
- Identification and Tracking
- Climate Forecasting
- Replacing entire models

Challenges with machine learning

- Labeling
  - Generating labeled data can be immensely time consuming
  - Bias in labeling remains an issue
- Interpretability(Black Box effect)
  - Even when ML works, we don't necessarily learn anything.
  - Danger of finding capable tools but not understanding how they work.
  - Often models are not testable
- Data quality
  - Bad data in = Bad Results out
- Worst Practices
  - Lack of validation sets, blind approaches, overfitting, etc.

Discussion Points

- As you hear presentations and during the discussion I'd like you to think about several points
  - What is the place of ML techniques for learning from modelling and observational datasets to improve model parameterizations.
  - Addressing ways in which ARM and ASR can improve their accessibility to ML solutions.
  - Are there techniques from the field that could help ARM/ASR address uncertainty quantification.
  - Are there any immediate needs the ARM/ASR community has that could benefit from machine learning?
  - Is there anything we as a community could better do to reach out to the broader ML community?

Pacific Northwest NATIONAL LABORATORY

Agenda

4:00-4:20	Joseph Hardin	Introduction & Overview of ARM ML Resources and plans
4:20-4:30	Jiali Wang	Fast domain aware neural network emulation of a planetary boundary layer parameterization in a numerical weather forecast model
4:30-4:40	Yangang Liu	Machine Learning for Cloud Microphysics Parameterization
4:40-4:50	Shuaiqi Tang	A Machine Learning Framework for ARM Data Quality Analysis: MWR Rain Contamination Detection
4:50-5:00	Ed Luke	Scanning radar sea-clutter mitigation using deep neural networks
5:00-5:10	Vanessa Przybylo	Ice Particle and Aggregate Simulator
5:10-5:20	Jitendra Kumar	Deep Convolutional Neural Networks for Hydrometeor Classification Using Dual Polarization Doppler Radars
5:20-5:30	V. Chandrasekar	AI and machine learning with weather radars
5:30-6:00		Discussion and Wrap-up

Discussion Points

- What is the place of ML techniques for learning from modelling and observational datasets to improve model parameterizations.
- Addressing ways in which ARM and ASR can improve their accessibility to ML solutions.
- Are there techniques from the field that could help ARM/ASR address uncertainty quantification.
- Are there any immediate needs the ARM/ASR community has that could benefit from machine learning?
- Is there anything we as a community could better do to reach out to the broader ML community?

- How do you feel the format of the session was? Right timing or fewer longer talks?

Pacific Northwest
NATIONAL LABORATORY

- Questions, concerns, follow up?
- Email joseph.hardin@pnnl.gov and I’d be happy to help or put you in contact with someone that can.