MACHINE LEARNING AND DATA SCIENCE FOR BIODIVERSITY

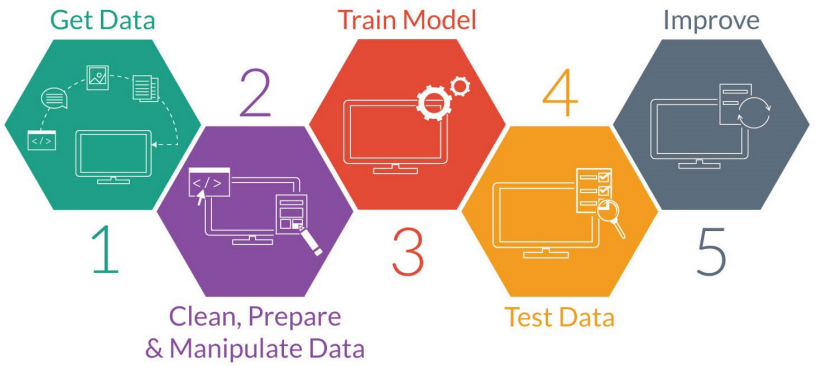
WILLIAM STEIMEL

ADVISOR – YUICHIRO MIYAMOTO

TABLE OF CONTENTS

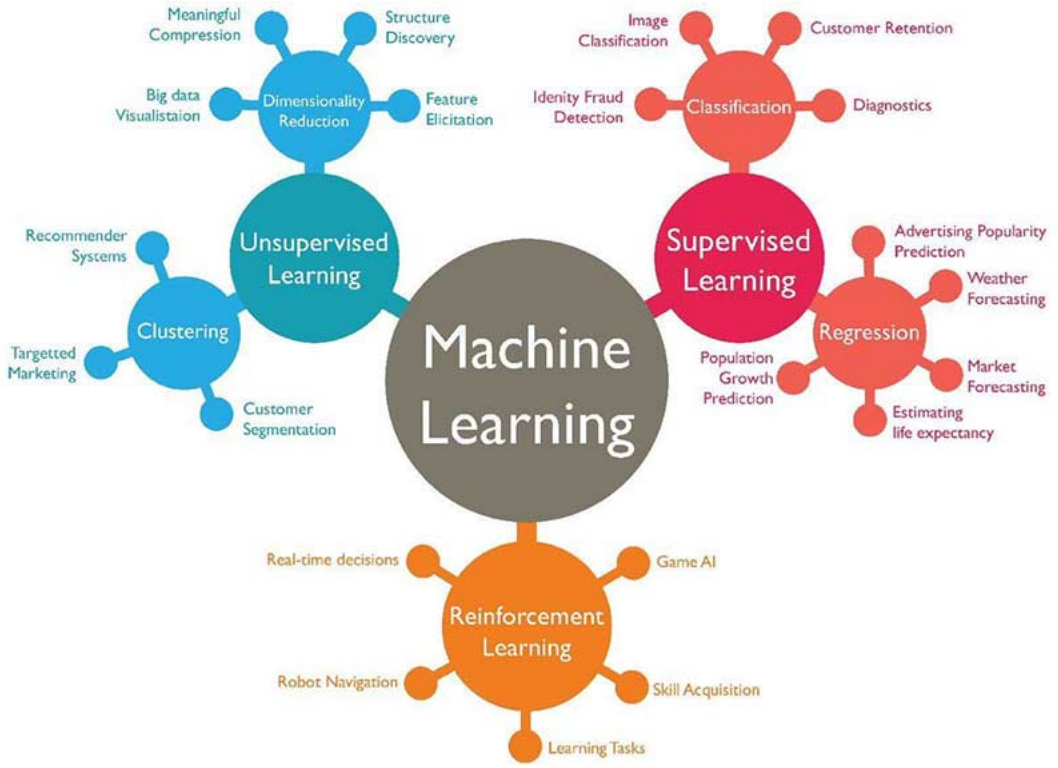
WHAT IS MACHINE LEARNING?

- Machine learning is a subset of the field of Artificial Intelligence
- Machine Learning - "Field of study that gives computers the ability to learn without being explicitly programmed" – Arthur Samuel 1959
- As we all know, more and more data each year is being generated through online services, sensors, e-mails, social media and other sources.
  - A machine learning algorithm learns to make decisions (predictions) based on data alone



WHAT AM I STUDYING?

- **Data Analysis/Applied Machine Learning**
- **Programming (R, Python, SQL)**
- **Supervised Learning (Classification/Regression)**
  - Linear Regression
  - Logistic Regression
  - Decision Trees
  - Support Vector Machines
  - Naïve Bayes
  - Random Forest
  - Neural Network
  - Etc.
- **Unsupervised Learning**
  - Clustering Methods
  - Dimensionality Reduction Methods
  - Etc.
- Essentially studying how to use each algorithm type and how I could use them in a business or applied setting
  - Also studying how to improve algorithm performance through methods like:
    - Hyperparameter Optimization (Improvement of algorithm performance through hyperparameter (knob) adjustment)
    - Feature Engineering (Creation of new feature data based on available data)



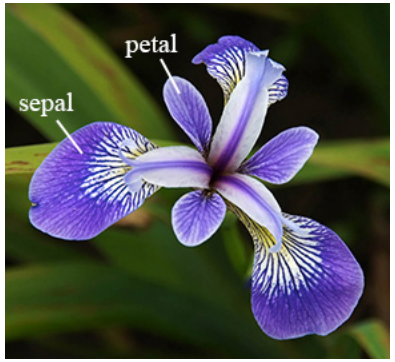
APPLICATIONS OF MACHINE LEARNING

- Some examples of how Machine Learning is being used in the real world:
  - Product Recommendations – Amazon Product Recommendations
  - Content Recommendations – Netflix/Youtube Recommendations based on viewing habits
  - Medical Research and Diagnosis – Detection of Pneumonia or Other diseases with X-Rays
  - Self Driving Vehicles – Tesla Autopilot
  - Fraud Detection – Banking Fraud Detection Systems
  - Search Engines – Google Search Engine Optimization
  - Social Media – Facebook/Instagram Feed
  - Many More

QUICK EXAMPLE OF MACHINE LEARNING

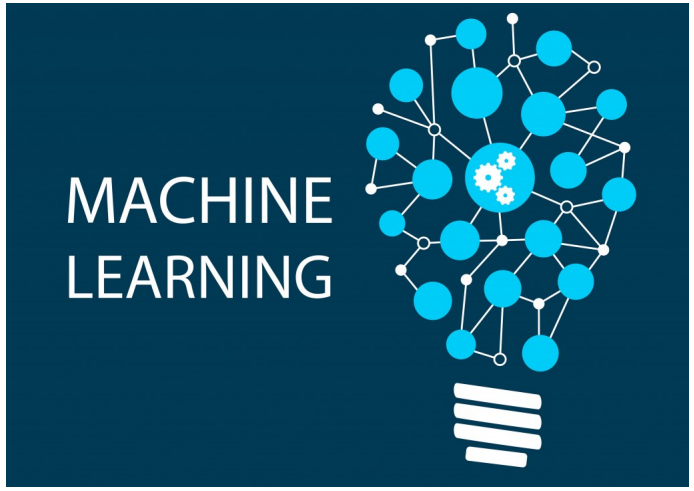
- **Iris Dataset Ronald Fisher – 1936**
  - Famous Dataset used for beginners in Machine Learning classification methods
  - **"THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS"**
- 4 Descriptive Features: (x_1, x_2, x_3, x_4)
  - Sepal Length
  - Sepal Width
  - Petal Length
  - Petal Width
- Target Variable: (y)
  - Species (Setosa, Versicolor, Virginica)
- Goal: Predict/Classify (Species) y based on feature values (Sepal Length x_1, Sepal Width x_2, Petal Length x_3, Petal Width x_4)
- Types of Data that Machine Learning can utilize:
  - Tabular
  - Images
  - Text
  - Audio
  - Anything that can be represented in a vector

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa



WHAT CAN MY PROJECT DO FOR BIODIVERSITY?

- It seems Machine Learning methods are already being applied to many Biodiversity related projects around the world.



WHAT CAN MY PROJECT DO FOR BIODIVERSITY?

- Computer Vision – A sub area of Machine Learning related to the processing of image and video data
  - Self Driving Cars
  - Facial Recognition
  - Image Classification (Is this a hotdog or not hotdog?)
- **Identifying Land Patterns from Satellite Imagery in Amazon Rainforest using Deep Learning**
  - Computer Vision Task (Image Data)
  - Many researchers have attempted to use Deep Learning to classify images and identify deforestation in the Amazon
  - One model achieved accuracy of 96.71 %
  - These models can be used to identify where deforestation is occurring and help policy makers track changes in land usage more effectively.

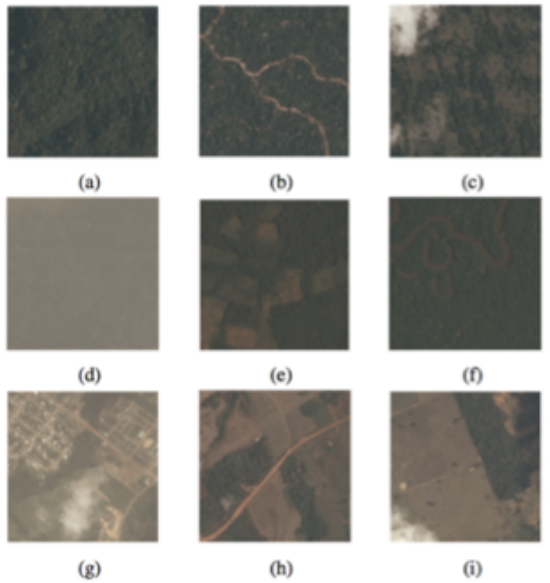


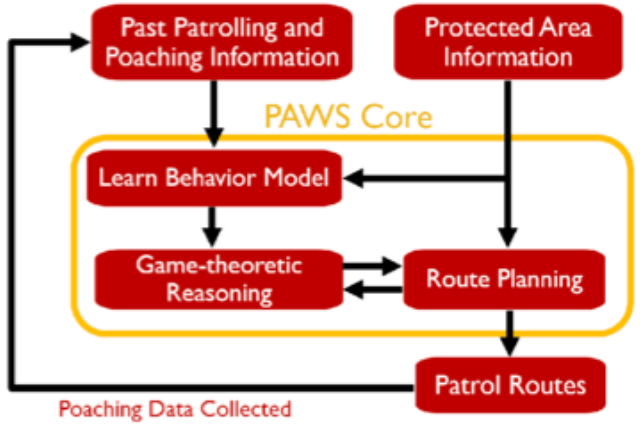
Fig. 3: Sample Chips and Their Labels (a) primary (b) roads + primary (c) partly_cloudy + primary (d) haze + primary (e) cultivation + primary (f) water + primary (g) habitation + partly_cloudy (h) agriculture + roads + primary (i) agriculture + pasture + primary + partly_cloudy

WHAT CAN MY PROJECT DO FOR BIODIVERSITY?

- **Deep Learning for Large Scale Biodiversity Monitoring (Bloomberg data for good 2015)**
  - ”Traditional Biodiversity monitoring involves periodically sending observers to a pre-determined set of survey sites to collect data over relatively short survey windows”
  - This paper attempts to solve this challenge by proposing a data driven approach to biodiversity monitoring that uses both machine learning and sensor data.
  - Types of Sensors used include:
    - Microphones (Audio)
    - Cameras (Image) – Visual, Thermal, IR, hyperspectral
    - Accelerometers (Speed- Numeric/Time Series)
  - Case Studies from this research:
    - Detecting Rare Species – Human surveys of rare species are expensive and automated monitoring of rare species was used at a cheaper more efficient rate.
      - Use Case- Finding Possible Rare Species that are hard to find by human survey.
    - Monitoring Populations through Time – Automated monitoring was able provide the required statistical power at a lower cost than traditionally in pilot cases at Great Barrier Reef Marine Park, and California Coastal National Monument.
      - Use Case- Analysis of Population trends over time
    - Detecting Invasive Species – Automated Sensors at Airports and Ports could be useful for detecting invasive species. This research team is currently working with the U.S Geological Survey on applying these sensors in Guam.
      - Use Case- Early Detection of Invasive Species

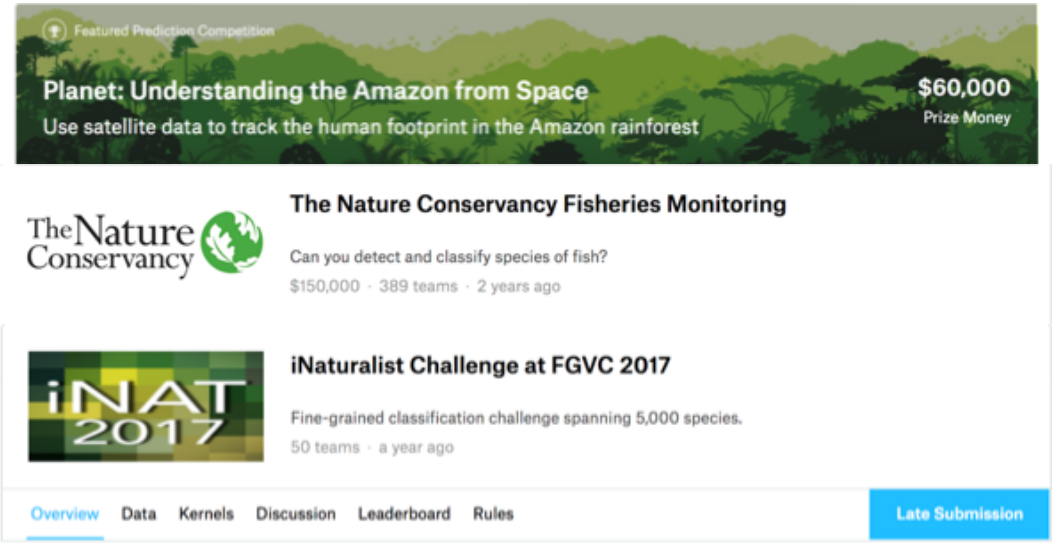
WHAT CAN MY PROJECT DO FOR BIODIVERSITY?

- PAWS – Protection Assistant for Wildlife Security (University of Southern California Viterbi: School of Engineering)
  - Input Data – (GPS(Geospatial) /crime data) information about previous patrols and poaching activities
  - Output – Uses Machine Learning (Ensemble of decision trees) to output potential patrol routes as output
  - Goal – Predict where poachers will strike next to protect already threatened species



WHAT CAN MY PROJECT DO FOR BIODIVERSITY?

- What is Kaggle?
  - A platform for competitive Data Science and Machine Learning
  - Data Scientists, Machine Learning Engineers, and Students around the world can compete for money and solve data related challenges.
- Biodiversity Related Challenges on Kaggle Recently



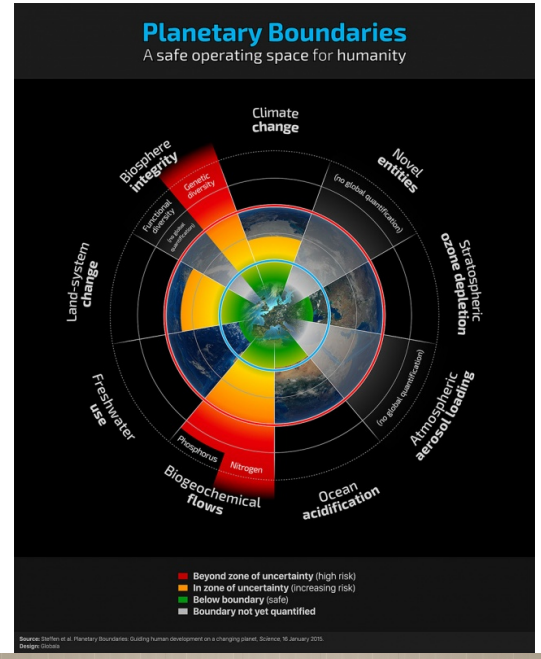
Contest Goal – Detect Deforestation in the Amazon with satellite image data. - Image Data (Previously Mentioned)

Contest Goal – Develop an algorithm that will automatically detect and classify fish caught by fishing boats. This model will assist with detection of illegal fishing that threatens marine ecosystems. – Image Data

Contest Goal – Sponsored by google with the aim of classifying over 5,089 animal and plant species through just images

WHAT DOES BIODIVERSITY MEAN TO MY PROJECT?

- According to the World Wide Fund for Nature we are losing around 10,000 species each year which is estimated to be about 1,000 to 10,000 higher than the natural extinction rate.
- **Global Planetary Boundaries Theory-** The Earth has 9 planetary boundaries which if exceeded will shift the Earth system to a point of no return. (Safe operating space for humanity)
  - Biosphere integrity/genetic diversity already in the red meaning we have surpassed our limits
- There are currently a lot of projects related to applying Machine Learning to Biodiversity but there are definitely more opportunities to apply Machine Learning in the Biodiversity realm.
  - Machine Learning has the opportunity to take our increasingly data driven world and create knowledgeable systems for biodiversity preservation and analysis.
  - Transdisciplinarity thinking is essential
  - As long as there is data the possibilities are endless!
- Maybe In the future, I will pursue more research related to Biodiversity in the near future as there is plentiful data in the area.
  - GBIF | Global Biodiversity Information Facility - Free and open access to biodiversity data
    - https://www.gbif.org/



THANK YOU !

SOURCES

- Machine Learning
  - https://www.ibm.com/developerworks/community/blogs/jfp/entry/What_Is_Machine_Learning?lang=en
  - Fisher,R.A. "The use of multiple measurements in taxonomic problems" Annual Eugenics, 7, Part II, 179-188 (1936); also in "Contributions to Mathematical Statistics" (John Wiley, NY, 1950).
- Papers
  - https://arxiv.org/pdf/1809.00340.pdf
  - http://bio.research.ucsc.edu/people/croll/pdf/Klein_2015.pdf
- Misc Projects
  - https://www.cais.usc.edu/projects/wildlife-security/
- Kaggle
  - https://www.kaggle.com/c/the-nature-conservancy-fisheries-monitoring
  - https://www.kaggle.com/c/planet-understanding-the-amazon-from-space
  - https://www.kaggle.com/c/inaturalist-challenge-at-fgvc-2017
- Biodiversity
  - http://wwf.panda.org/our_work/biodiversity/biodiversity/
  - http://wwf.panda.org/our_work/biodiversity/biodiversity_and_you/
  - http://www.futureearth.org/blog/2015-feb-6/planetary-boundary-biodiversity