

Lecture Notes for **Machine Learning in Python**

Professor Eric Larson
Introduction, Syllabus, Data Types

Class Logistics and Agenda

- Agenda:
 - Introductions
 - Syllabus and Course Overview
 - What is Machine Learning?
 - Types of Data
 - Numpy/Pandas Demo
- My approach to this course:
 - Programming
 - Math
 - **Applications** and **Analytics**

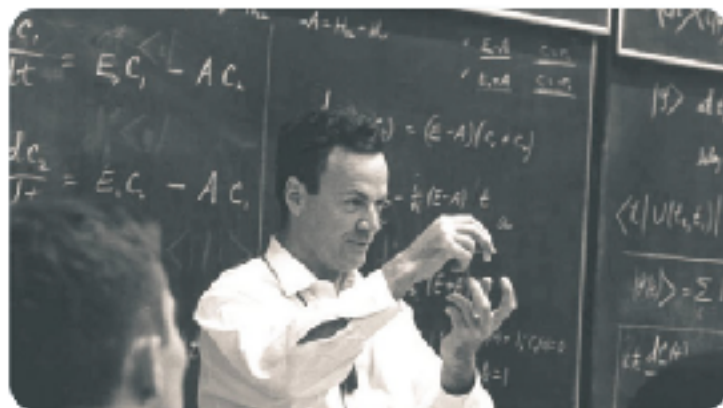
Introductions & Course Syllabus



Richard Feynman @ProfFeynman · 12h

Don't just teach your students to read.

- Teach them to **question** what they read, what they study.
- Teach them to **doubt**.
- Teach them to **think**.
- Teach them to make mistakes and learn from them.
- Teach them how to understand something.
- Teach them how to teach others.



Richard Feynman @ProfFeynman · 21h

You cannot get educated by this self-propagating system in which people study to pass exams, and teach others to pass exams, but nobody knows anything.

You learn something by doing it yourself, by asking questions, by thinking, and by experimenting. 🧠



Introductions

- Me
 - Eric 👍
 - Dr. Larson 👍
 - Prof. Larson 👍
 - Other 👎
- You
 - Name
 - Where you grew up
 - Department
 - Grad/Undergrad
 - Something true or false

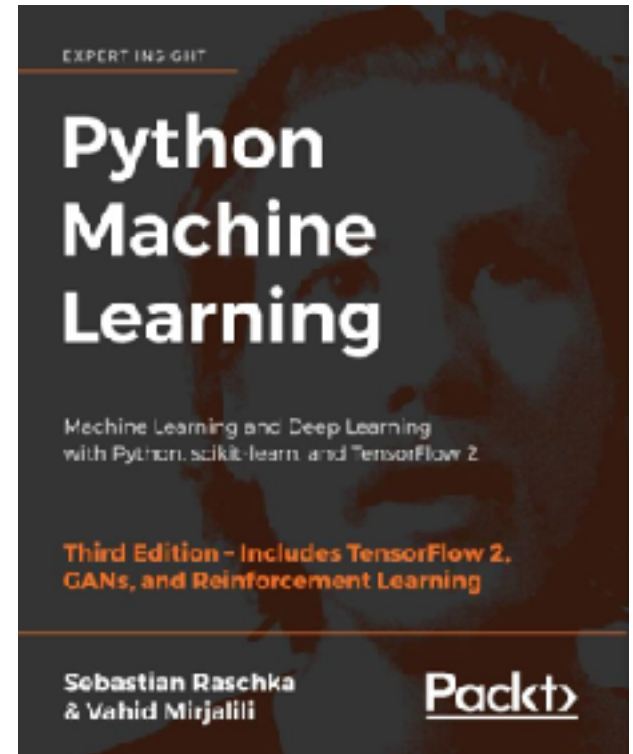
Choices:

High School
College Job
ML Instructor
Marvel
Kids

Limited Introduction because of Hybrid Class

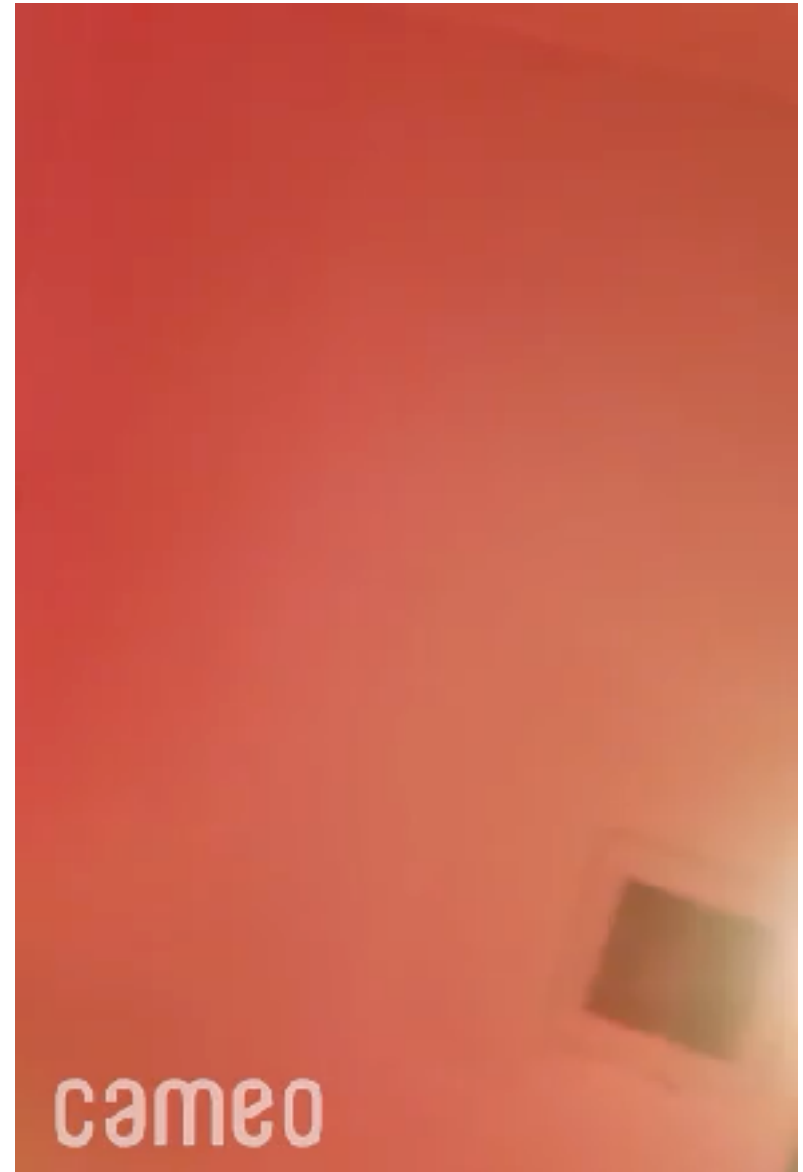
FAQ

- Text:
 - **Recommended:** Python Machine Learning, Raschka & Mirjalili, Second edition
- Use Canvas for posted course material
- Prerequisites:
 - Linear algebra, calculus (multivariate)
 - Basic statistics and probability
 - Basic python programming
- Version of python: 3.X
 - Install through Anaconda
 - Use conda environments
 - JupyterLab (or notebook)
- Most Used Libraries: Numpy, Pandas, Scikit-Learn, Matplotlib, Keras with Tensorflow



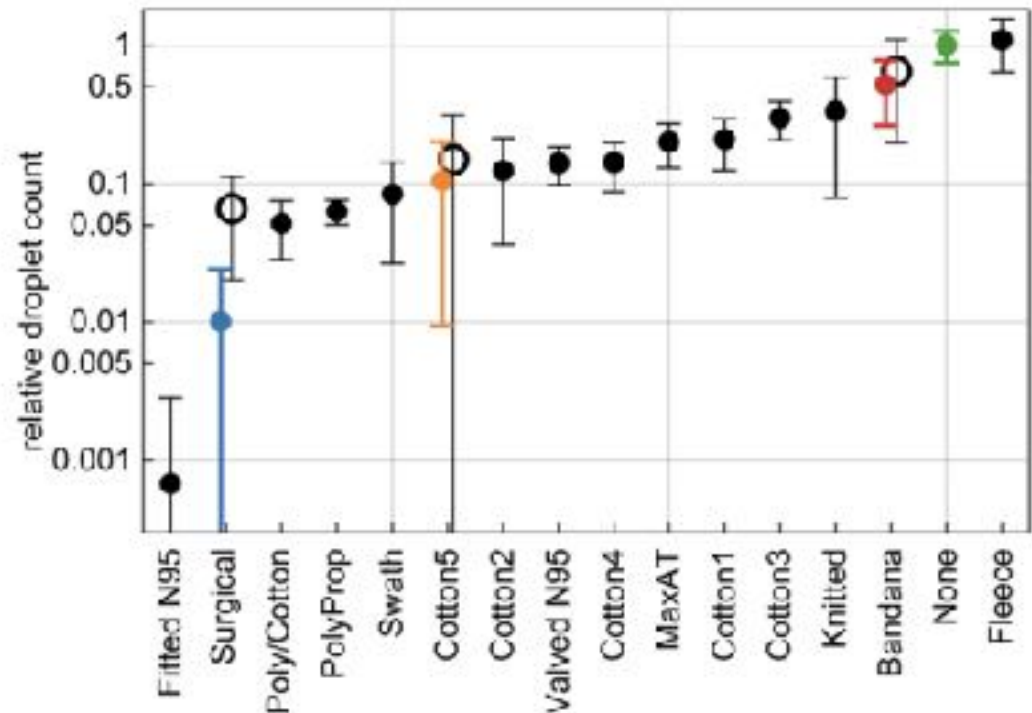
Canvas Syllabus

- Lab Assignments
- Grading Rubrics
- Participation
- Course Schedule
- Flipped Assignments
- Difference between 5000 and 7000



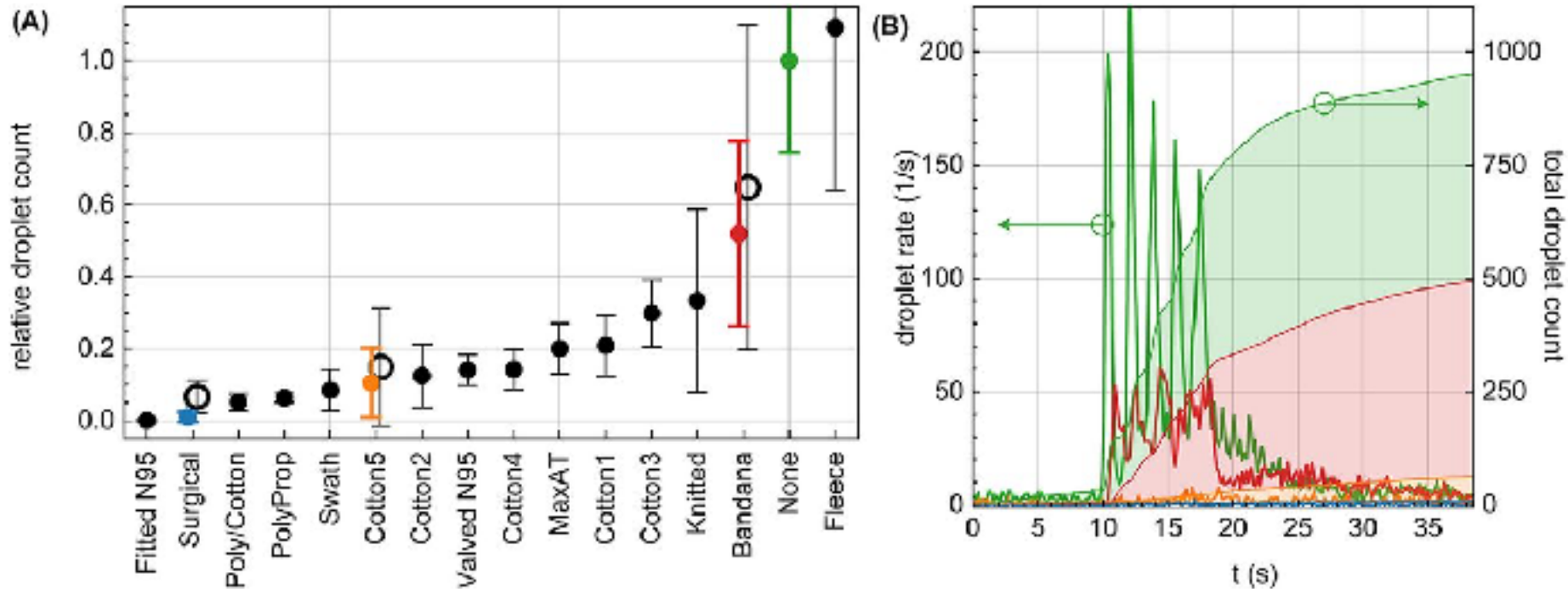
Other topics

- Due to the pandemic: Participation is graded via the quizzes!
- Other topics specific to pandemic:
 - Attendance (ingress/egress, wipe down surfaces, do NOT congregate at doors)
 - Zoom etiquette (chat/video)
 - Flipped assignment
 - Social distancing

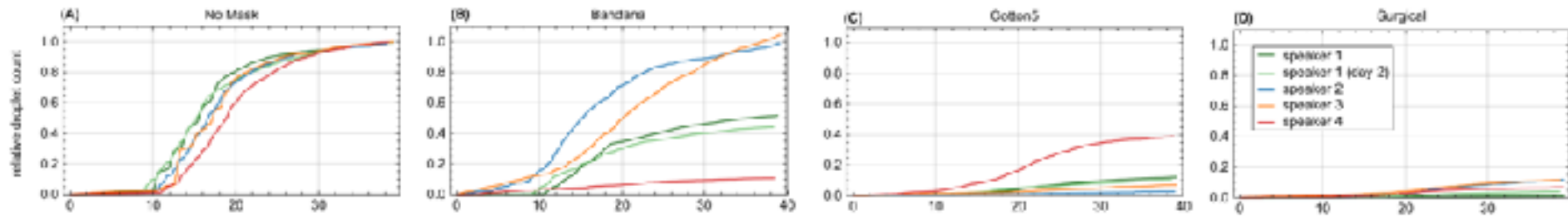


<https://advances.sciencemag.org/content/early/2020/08/07/sciadv.abd3083>

Another view of the data



only deals with droplet transmission, not aerosol



<https://advances.sciencemag.org/content/early/2020/08/07/sciadv.abd3083>

Class Overview, by assignment

- **Lab One:** Visualize data and extract some features
- **Lab Two:** Analyze Images, Use dimensionality Reduction
- **Lab Three:** Program Logistic Regression in style of Sci-kit Learn
- **Lab Four:** Program NN Back propagation from Scratch, implement Adaptive Gradient Techniques
 - Use given dataset for this lab
- **Lab Five:** wide and deep networks
- **Lab Six:** Classify Images with Convolutional Networks
- **Lab Seven:** Classify Text with Recurrent Neural Networks

All Assignments posted on Canvas, with Rubric
Everything is a team assignment except quizzes

Class Overview, by topic

Table Data
Visualization

Numpy, Pandas, Seaborn
Overviews with some in-depth discussion

Dimension
Reduction and
Image Processing

Scikit-learn, Scikit Image,
Intuition only, Some mathematics

Linear and
Logistic
Regression

Numpy, Recreate API for Scikit-learn
Detailed mathematics for simple optimization
intuition for advanced optimization

Neural Networks
and Back Prop.

Numpy
Detailed mathematics for NN operations

Wide and Deep
Networks

Convolutional
Networks

Recurrent
Networks

Keras, Tensorflow
Intuition, Detailed implement.

Ethics in
Language Models

ConceptNet
Case studies

Is this plagiarism in this class?

- Copying code/text from another source without citing it
 - A. Yes, plagiarism!
 - B. No, its fine!
- Copying code/text from another source, citing at the end of the assignment in a blanket statment (but not making it clear which part of the assignment was from another source)?
 - A. Yes, plagiarism!
 - B. No, its fine!
- Copying code, citing the source directly next to the code, and commenting on what parts were changed?
 - A. Yes, plagiarism!
 - B. No, its fine!
- Copying text directly and citing the source with the text, but not placing the text in quotes.
 - A. Yes, plagiarism!
 - B. No, its fine!

Machine Learning Overview



What is Machine Learning?

Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. **Machine learning** focuses on the development of computer programs that can change when exposed to new data.

What is machine learning? - Definition from WhatIs.com
[whatis.techtarget.com/definition/machine-learning](https://www.whatis.techtarget.com/definition/machine-learning)

About this result • Feedback

- **Beware of this definition:**
 - full of imprecise, loaded words:
 - intelligence, learning
 - ignores social structures, ethics, deployment, and that all results are interpreted by a human

Machine Learning

One Small Piece of Artificial Intelligence

Data Mining

ML

Prediction Methods

- Use some variables to predict unknown or future values of other variables

Description Methods

- Find human-interpretable patterns that describe the data.

ML

- Classification
- Regression
- Deviation Detection
- Clustering
- Association Rule Discovery
- Sequential Pattern Discovery

Problem Types in Machine Learning





kaggle

Customer Solutions

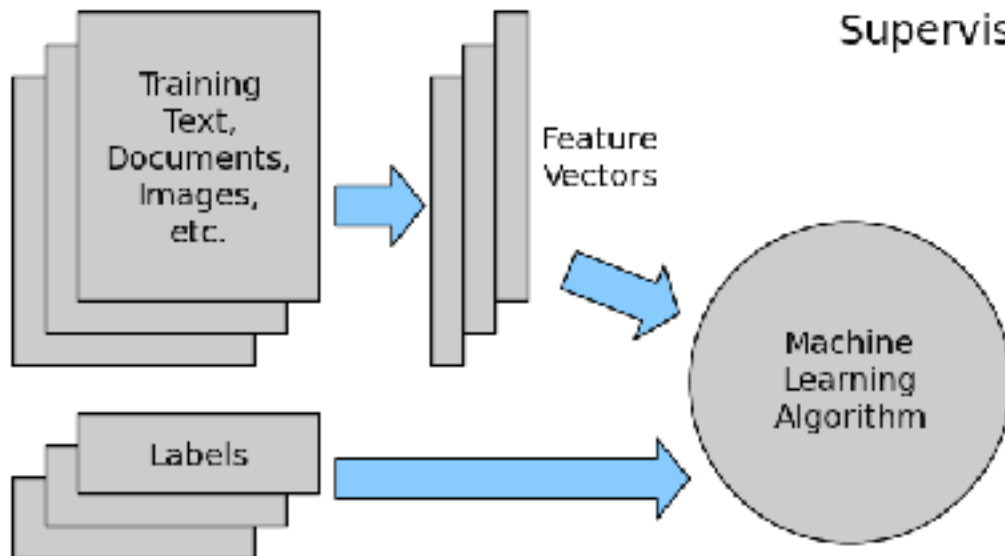
Competitions

Community ▾

Active Competitions

		Click-Through Rate Prediction Predict whether a mobile ad will be clicked	21 days 1512 teams \$15,000
		National Data Science Bowl Predict ocean health, one plankton at a time	56 days 430 teams \$175,000
		Driver Telematics Analysis Use telematic data to identify a driver signature	56 days 686 teams \$30,000

Classification and Regression



Supervised Learning Model

- *Training* Instances: Features + Labels
- Find a *model* mapping class from values of features.
- Goal: Assign class to previously unseen instances

Example Classification: Malware

- Classify files as malware based on size and naming.
- Approach:
 - ♦ Use already classified malware files
 - ♦ Must translate name to set of features
 - ♦ **{malware, not malware}** decision forms the **class attribute**
 - ♦ Collect various malware examples and a number of safe files, providing labels for each and a set of features

Training Set

TID	Name	Size	Class
1	erte.dll	916 b	not
2	fufu.bin	1M	yes
3	exe.exe	1G	not
4	ex.py	113 b	not

Unknown

<i>TID</i>	<i>Name</i>	<i>Size</i>
1	asdf.dll	11b

Example Regression: Housing Price

- Predict a value of a given *continuous valued* variable based on the values of other variables
- Examples:
 - Predicting sales amounts of new product based on advertising expenditure.
 - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
 - Predicting House Sales

Training Set

<i>TI</i>	<i># Rms</i>	<i>Sq Ft</i>	<i>Zip</i>	<i>Price</i>
1	2	1125	74012	150K
2	2	2525	75155	200k
3	10	4678	90210	3M
4	4	2678	75154	350k

Unknown

<i>TI</i>	<i># Rms</i>	<i>Sq Ft</i>	<i>Zip</i>
1	2	2200	75115

Example Classifying: Objects in Images

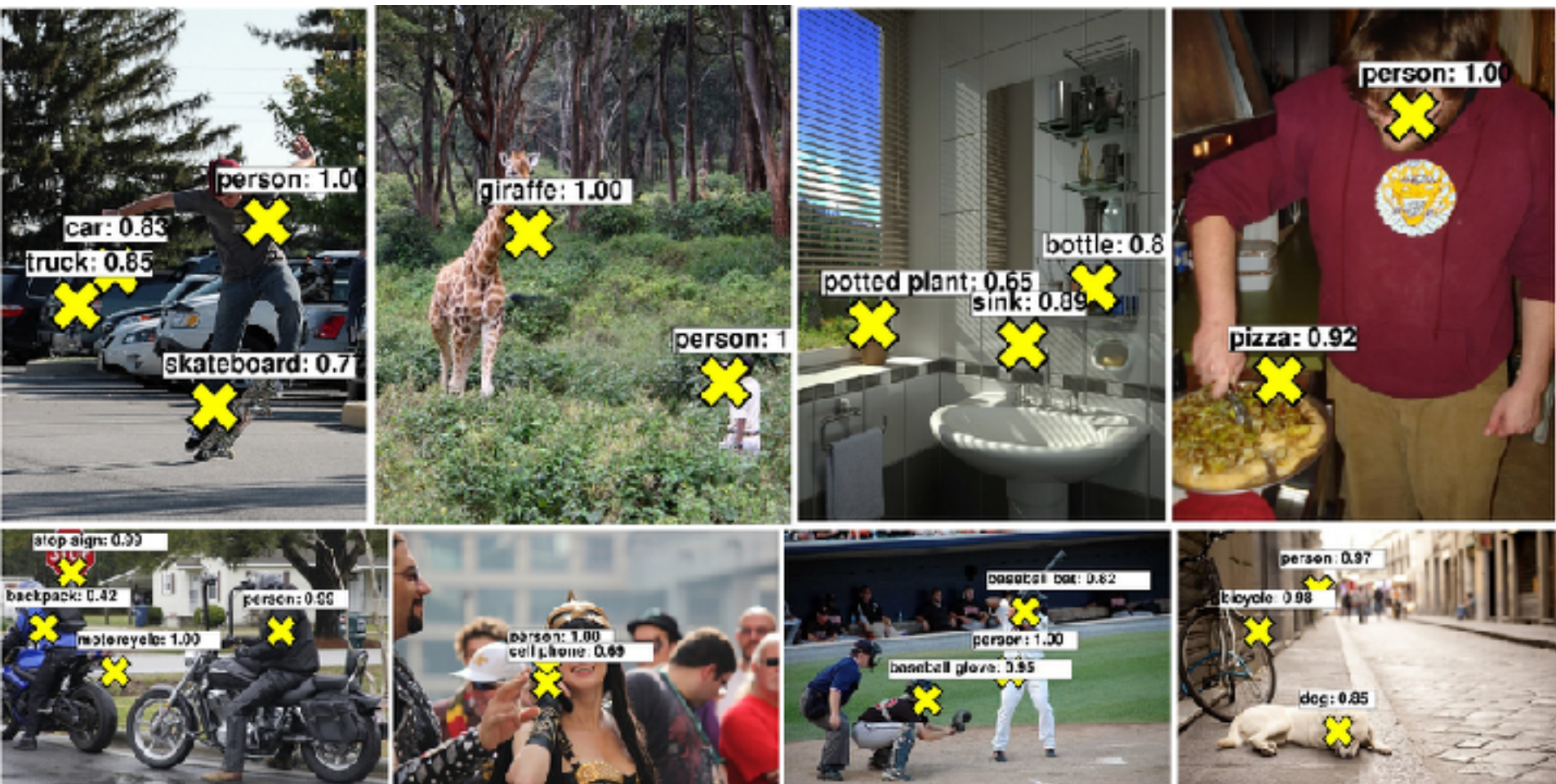


Image Net:

- 14 million images
- 200 Labeled Categories
- 1000 Location Labels

Attributes:

- Images

Self Test

- **A. Classification**
B. Regression
C. Not Machine Learning
- Dividing up customers by potential profitability?
- Extracting frequency of sound?

Me after that one free course I found from an add I clicked on.

```
from math import sqrt, ceil, floor, gcd
import numpy as np
import pandas as pd
import tensorflow as tf
import keras
import matplotlib.pyplot as plt
import nltk
import scipy
import click
import collections
import cryptography, crypt, astropy, csv, cython, curl, cv2, math, cmath
import gensim
import http
import seaborn
import sklearn

# 10:20 PM
print("Hello world!")
```



Types of Data and Categorization

Table Data

- **Table Data:** Collection of data **instances** and their **features**
- **Python:** Pandas Dataframe
- **R:** Data.frame
- **Matlab:** Table Class
- **C++:** Trick Question
make it yourself

Attributes, columns,
variables, fields,
characteristics, **Features**

Objects,
records,
rows
points,
samples,
cases,
entities,
Instances

<i>TID</i>	<i>Pregnant</i>	<i>BMI</i>	<i>Age</i>	<i>Diabetes</i>
1	Y	33.6	41-50	positive
2	N	26.6	31-40	negative
3	Y	23.3	31-40	positive
4	N	28.1	21-30	negative
5	N	43.1	31-40	positive
6	Y	25.6	21-30	negative
7	Y	31.0	21-30	positive
8	Y	35.3	21-30	negative
9	N	30.5	51-60	positive
10	Y	37.6	51-60	positive



Feature Type Representation

	Attribute	Representation Transformation	Comments
Discrete	Nominal	Permutation of values only. one hot encoding or hash function	If all employee ID numbers were reassigned, would it make any difference?
	Ordinal	Order must be preserved $\text{new_value} = f(\text{old_value})$ where f is a monotonic function. integer	An attribute encompassing the notion of good, better best can be represented equally well by the values {1, 2, 3} or by {0.5, 1, 10}.
Continuous	Interval	$\text{new_value} = f(\text{old_value}) + b$ f is monotonic through origin float	Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
	Ratio	$\text{new_value} = f(\text{old_value})$ f is monotonic through origin float	Length can be measured in meters or feet, but zero is zero

from Tan et al. Introduction to Data Mining

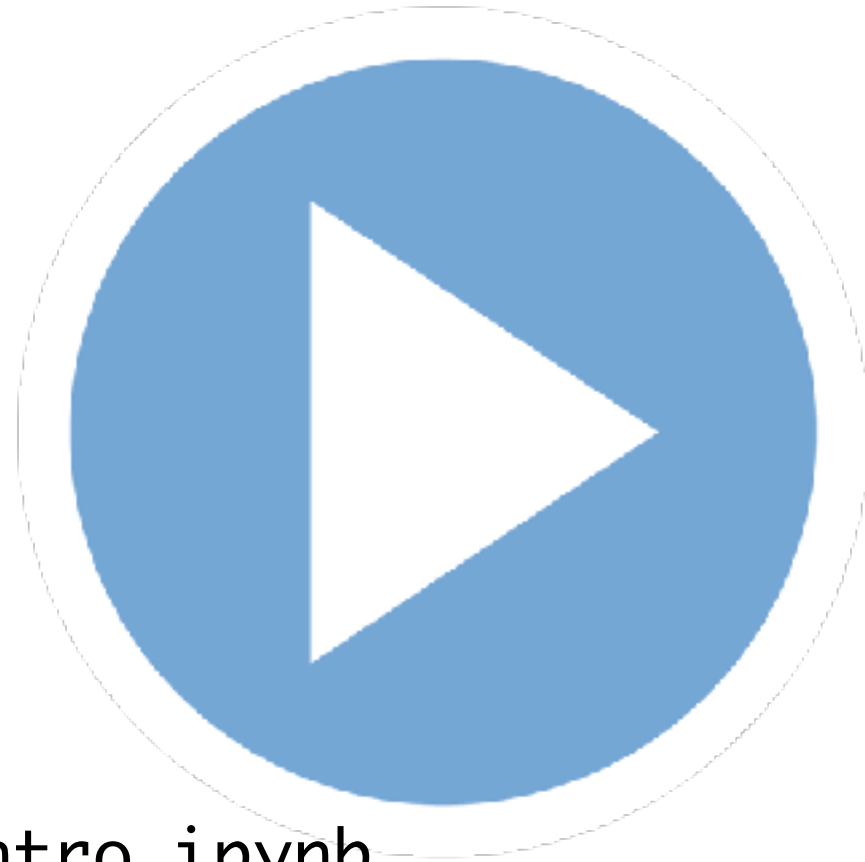
Self Test

- Are these **A. ordinal, B. nominal, or C. binary**?
 - military rank
 - ordinal
 - coat check number
 - nominal

Before Next Lecture

- Before next class:
 - install python on your laptop
 - install anaconda distribution of python
- Look at Python primer if you need review
 - I made ~4 hours of YouTube content...
 - <https://www.youtube.com/playlist?list=PL7IPdRN5E0YKCnVI-fvx8jOOCWVeGTsrV>

**If time:
Jupyter Notebooks**



`01_Numpy and Pandas Intro.ipynb`