

Lecture Notes for **Machine Learning in Python**



Professor Eric Larson
Introduction, Syllabus, Data Types

Class Logistics and Agenda

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

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

Sequential
Networks

Keras, Tensorflow
Intuition, Detailed implement.

Ethics in
Language Models

ConceptNet
Case studies

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 Sequential Networks

All Assignments posted on Canvas, with Rubric
Everything is a team assignment except quizzes, participation
You CANNOT makeup late quizzes, participation

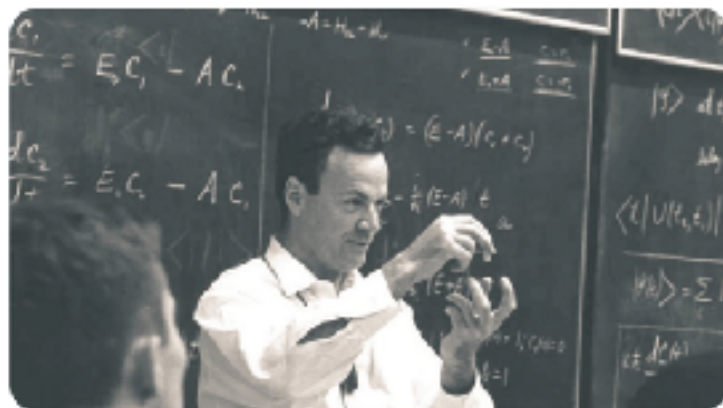
Introductions & Course Syllabus



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. 🧠



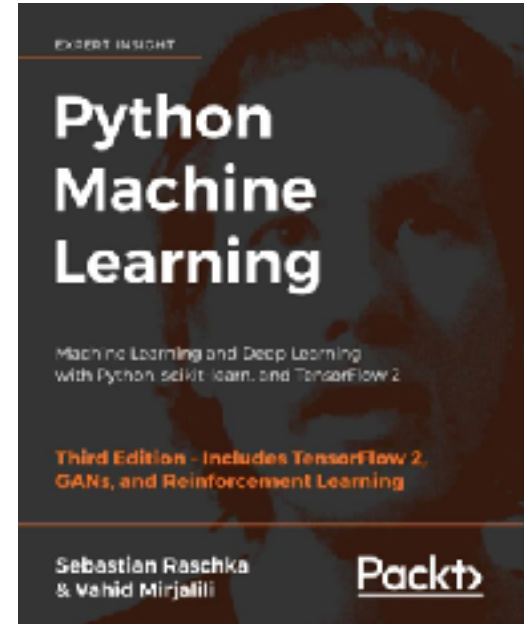
Introductions

- Me
 - Dr. Larson 👍
 - Prof. Larson 👍
 - PhD students: (Eric) 👁👁
 - Other 👎
- You
 - Name Department
 - Grad/Undergrad
 - Something true or false

Limited Introduction because of Class Size

FAQ

- Text:
 - **Recommended:** Python Machine Learning, Raschka & Mirjalili, Third Edition
- Use Canvas for posted course material
- Prerequisites:
 - Linear algebra & calculus (multivariate)
 - Basic statistics and probability
 - Basic OO programming, some python
- Version of **python: 3.X**
 - Install through **Anaconda** and pip
 - Use **conda** environments
 - JupyterLab (or **notebook**)
- Most Used Libraries: Numpy, Pandas, Scikit-Learn, Matplotlib, Seaborn, Tensorflow
- Use OIT Data Science Workshops



Canvas Syllabus

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

How will participation be graded?

- Participation will be graded in the course:
 - **Distance students** will answer these questions via **canvas upload** (same for Zoom)
 - upload “over” the last submission
 - must upload the questions each week for full credit
- In Class Students:
 - Live question answering (mostly attendance):
 - Do you think this will work?
 - A: **Yes** this is going to work.
 - B: This is **not** going to work.
 - C: My name was not on my card.
 - D: I (will/did) add an Alias to my card.

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 statement (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!

Is this plagiarism in this class?

- Using ChatGPT or other LLM that generates text/code/responses?
 - A. Yes, plagiarism!
 - B. No, its fine!
 - C. It might be okay, but students should:
 - 1) acknowledge when using it,
 - 2) add comments to code used to indicate our knowledge of the subject matter,
 - 3) check the accuracy and reliability of the output,
 - 4) do not use text word for word, only as an outline or exemplar of a possible answer

**Don't use a LLM at the detriment of your own understanding.
Don't use a LLM because your are unsure of your own understanding**

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
- **My definition:** a way to optimize model parameters for recognizing complex patterns in data

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

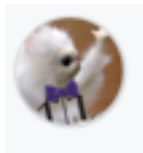
Problem Types in Machine Learning

- Inputs

- Outputs



1.23
-0.4
...



This is a repository for my
experience in Python and
purpose.



- Categories

- Numeric Data

- Images

- Free Text

- Times Series

classification

regression

image generation

text generation

auto encoding

- Categories

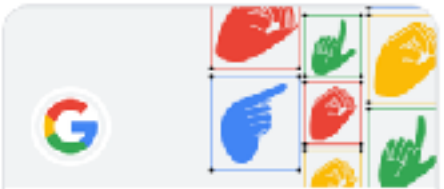






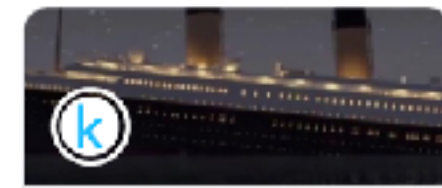
- Numeric Data

- Images

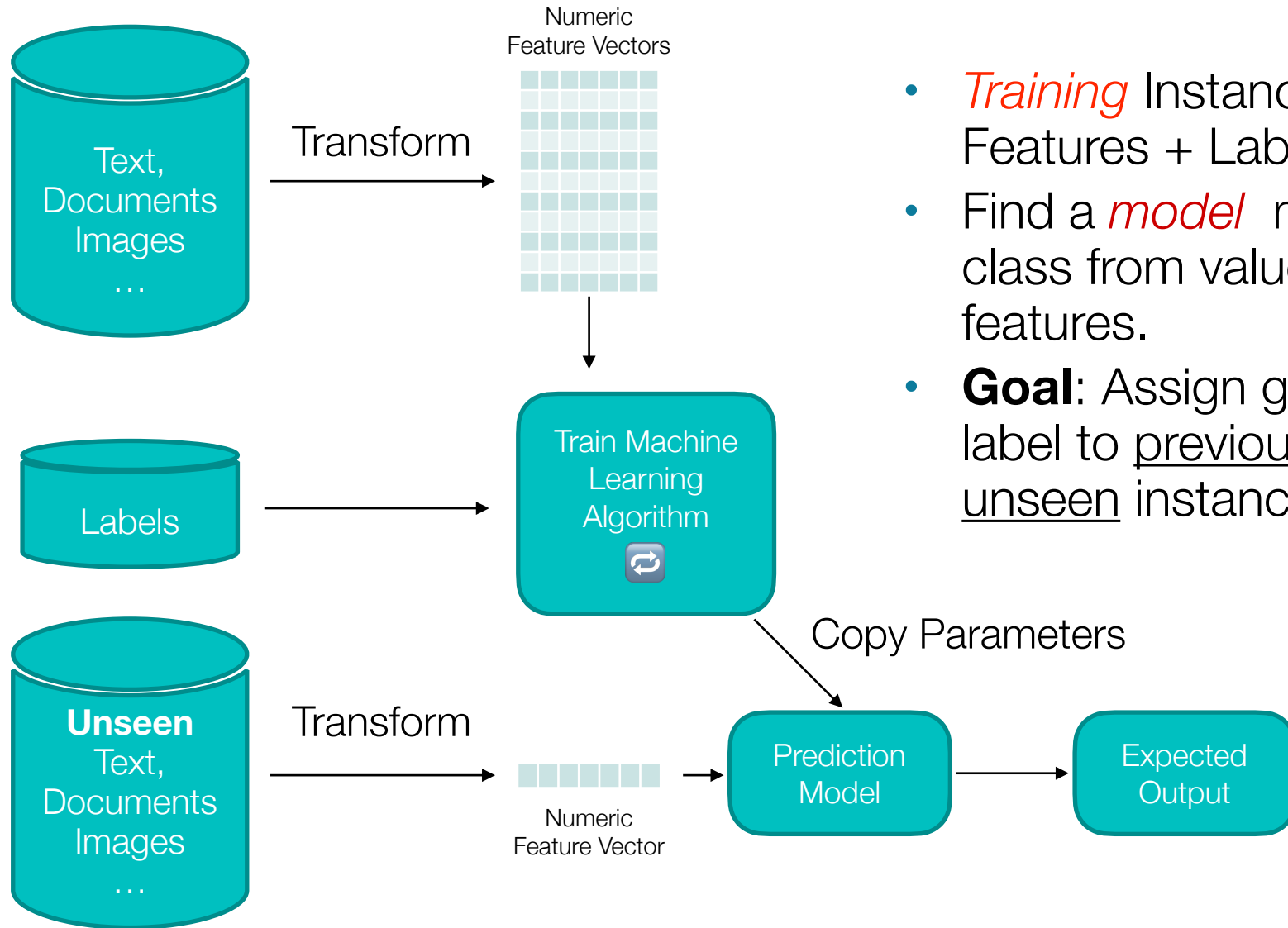
- Free Text

- Time Series

Problem Types in Machine Learning

 <p>Google - American Sign Language Fingerspelling...</p> <p>Train fast and accurate American Sign...</p> <p>Research · Code Competition</p> <p>1268 Teams</p> <p>\$200,000 3 days to go</p>	 <p>CommonLit - Evaluate Student Summaries</p> <p>Automatically assess summaries writt...</p> <p>Featured · Code Competition</p> <p>925 Teams</p> <p>\$60,000 2 months to go</p>	 <p>Bengali.AI Speech Recognition</p> <p>Recognize Bengali speech from out-of...</p> <p>Research · Code Competition</p> <p>317 Teams</p> <p>\$53,000 2 months to go</p>	 <p>CAFA 5 Protein Function Prediction</p> <p>Predict the biological function of a pro...</p> <p>Research · Code Competition</p> <p>1655 Teams</p> <p>\$50,000 10 hours to go</p>
 <p>Kaggle - LLM Science Exam</p> <p>Use LLMs to answer difficult science ...</p> <p>Featured · Code Competition</p> <p>1471 Teams</p> <p>\$50,000 2 months to go</p>	 <p>RSNA 2023 Abdominal Trauma Detection</p> <p>Detect and classify traumatic abdomi...</p> <p>Featured · Code Competition</p> <p>333 Teams</p> <p>\$50,000 2 months to go</p>	 <p>Predict CO2 Emissions in Rwanda</p> <p>Playground Series · Season 3, Episod...</p> <p>Playground</p> <p>1401 Teams</p> <p>Swag 10 hours to go</p>	 <p>Titanic - Machine Learning from Disaster</p> <p>Start here! Predict survival on the Tita...</p> <p>Getting Started</p> <p>14897 Teams</p> <p>Knowledge Ongoing</p>

Classification and Regression, Supervised



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

Some Popular Datasets

ImageNet



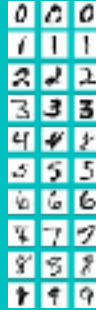
1M+

224 x 224 Color Image



1000 Classes
(prominent object)

MNIST



60k

28 x 28 Grey Image



10 Classes (digits)

Adult

#	feature	original feature
1	age	
2	workclass	
3	final_weight	
4	education	
5	edu_num	
6	marital_status	
7	occupation	
8	relationship	
9	race	
10	sex	
11	capital_gain	
12	capital_loss	
13	hours_in_week	
14	country	

5k

Census Demographics



Binary (salary > 50k?)

CoCo



200k Images

Large, Multi-sized Images



Location, Size, 80 Objects

Boston Housing

House/Neighborhood
Descriptions



House Price
\$\$

500 Examples

Translation



Language A



Language B

Many datasets

SQuAD



Question



Answer

100k+

Imdb



Movie/Actors/Director/+



Critic/Audience rating

50k reviews

Self Test

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