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

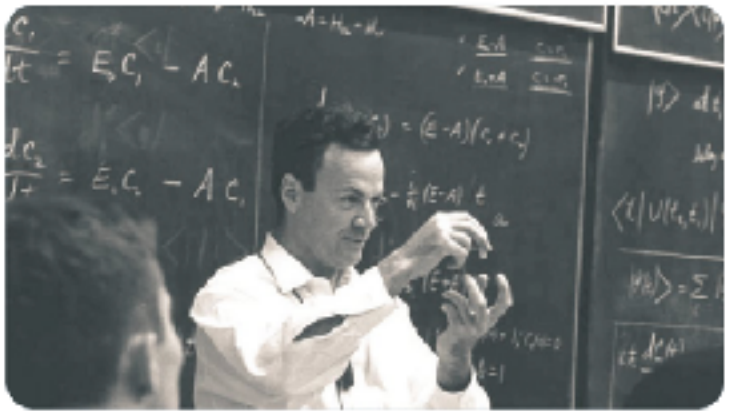
- My approach to this course:
  - Programming
  - Math
  - **Applications** and **Analytics**

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

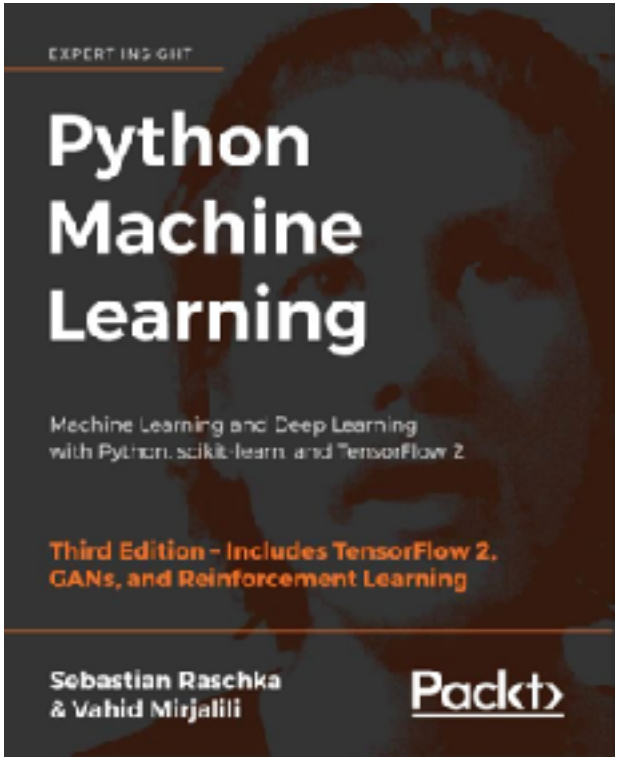
- Me
  - Eric 👍
  - Dr. Larson 👍
  - Prof. Larson 👍
  - Other 👎
- You
  - Name
  - Where you grew up
  - Department
  - Grad/Undergrad
  - Something true or false

Choices:
High School
College Job
ML Instructor
Marvel
Kids

Limited Introduction because of Hybrid Class

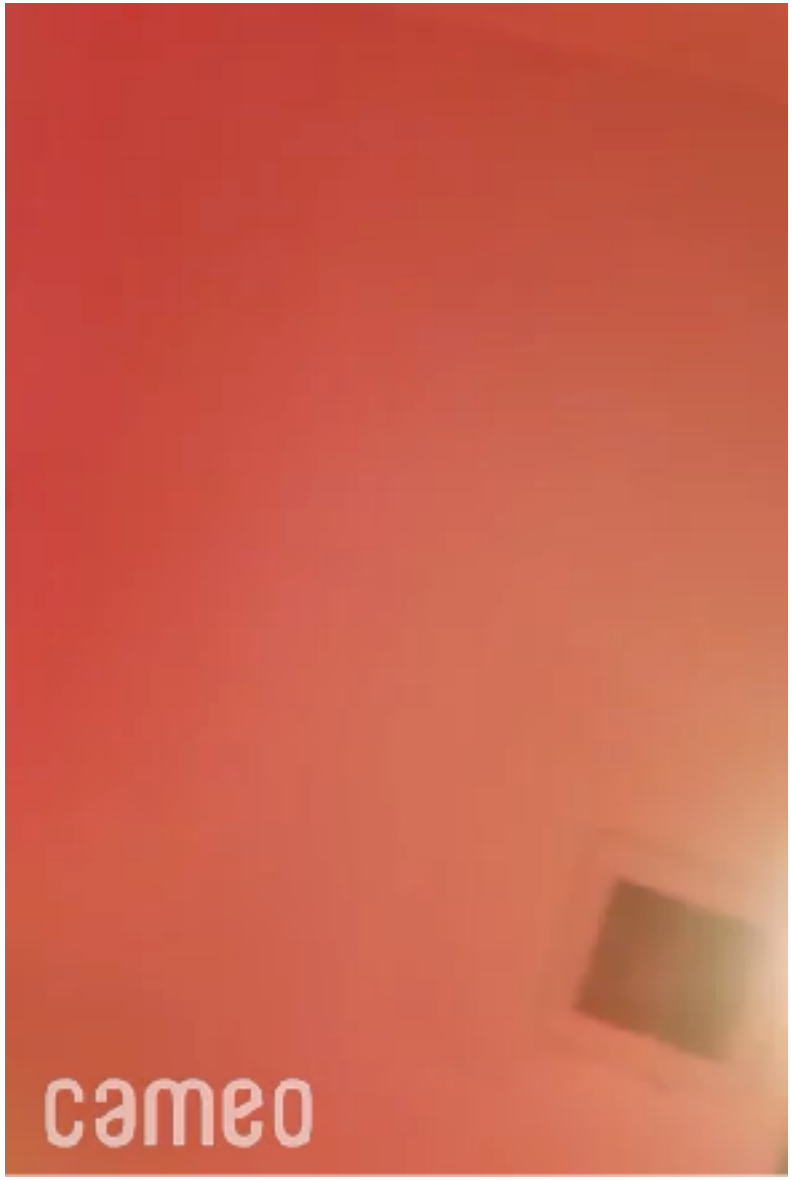
FAQ

- Text:
  - **Recommended**: Python Machine Learning, Raschka & Mirjalili, Third 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, Seaborn, Tensorflow



Canvas Syllabus

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



Pandemic topics

- Participation is graded via the quizzes! **Not lecture questions**
- Attendance (if you want, but wear a mask)
  - Zoom always available, and videos posted if you cannot make it to class
  - Zoom etiquette (chat/video)
- Flipped assignments (distance, in-class, team, etc.)

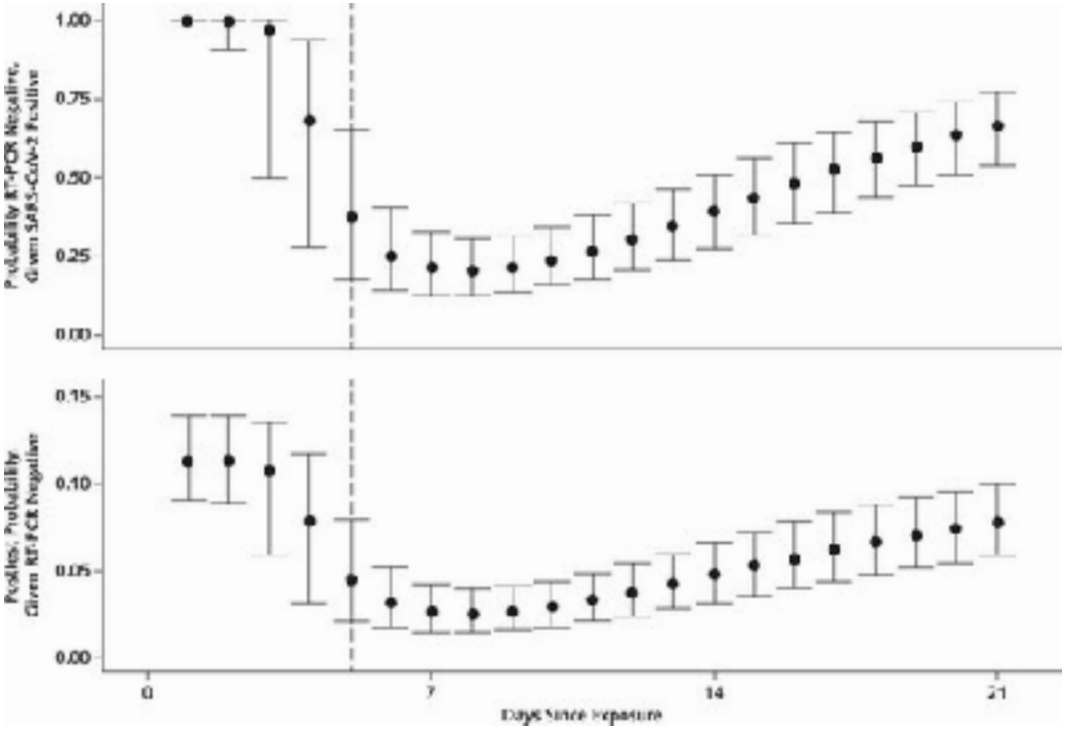
Over the 4 days of infection before the typical time of symptom onset (day 5), the probability of a false negative result in an infected person decreases from 100% (95% CI, 100% to 100%) on day 1 to 67% (CI, 27% to 94%) on day 4, although there is considerable uncertainty in these numbers. On the day of symptom onset, the median false-negative rate was 38% (CI, 18% to 65%) (Figure 2, *top*). This decreased to 20% (CI, 12% to 30%) on day 8 (3 days after symptom onset) then began to increase again, from 21% (CI, 13% to 31%) on day 9 to 66% (CI, 54% to 77%) on day 21.

Original Research | 18 August 2020

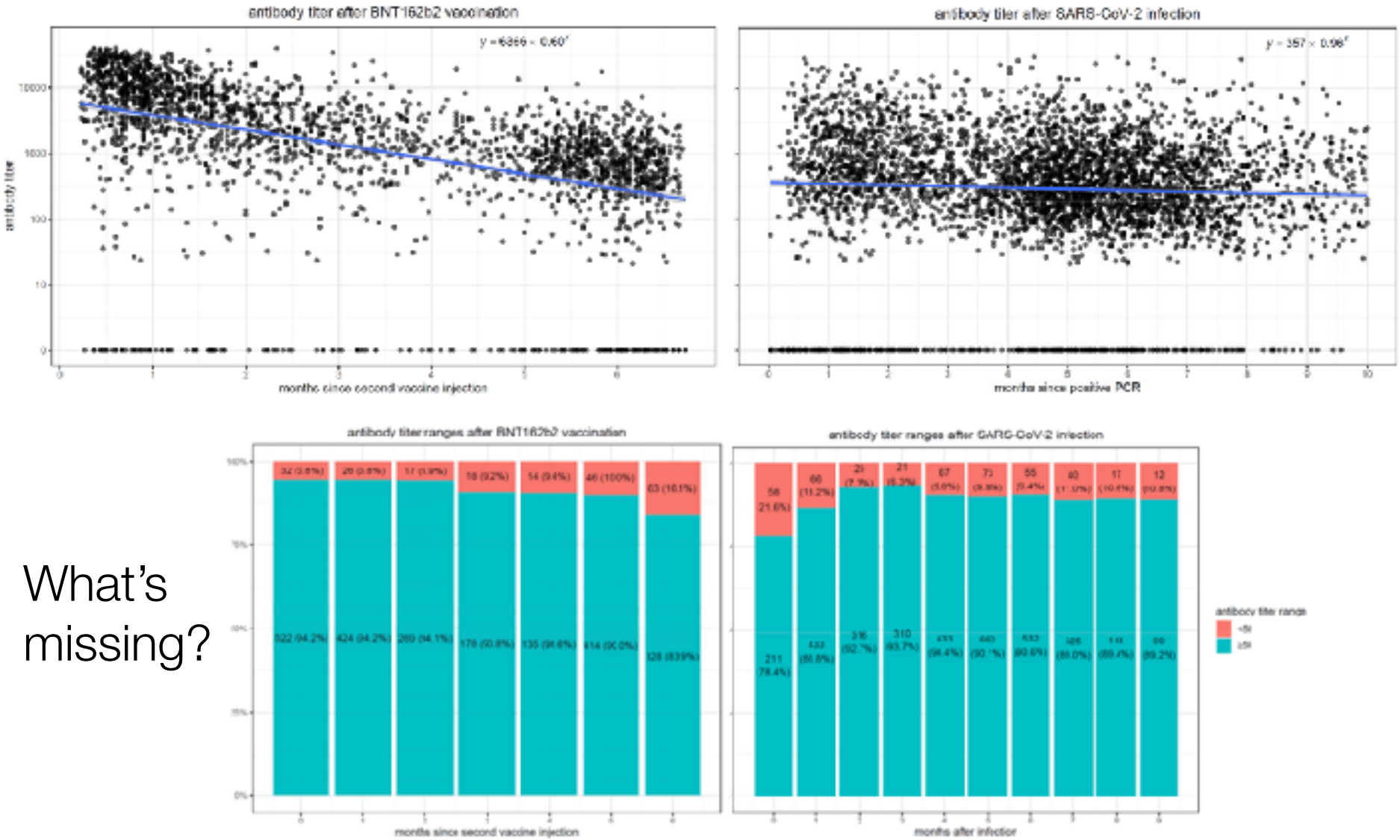
Variation in False-Negative Rate of Reverse Transcriptase Polymerase Chain Reaction–Based SARS-CoV-2 Tests by Time Since Exposure max

Author, Article and Disclosure Information

https://doi.org/10.7326/M20-1495 Eligible for CME Point of Care



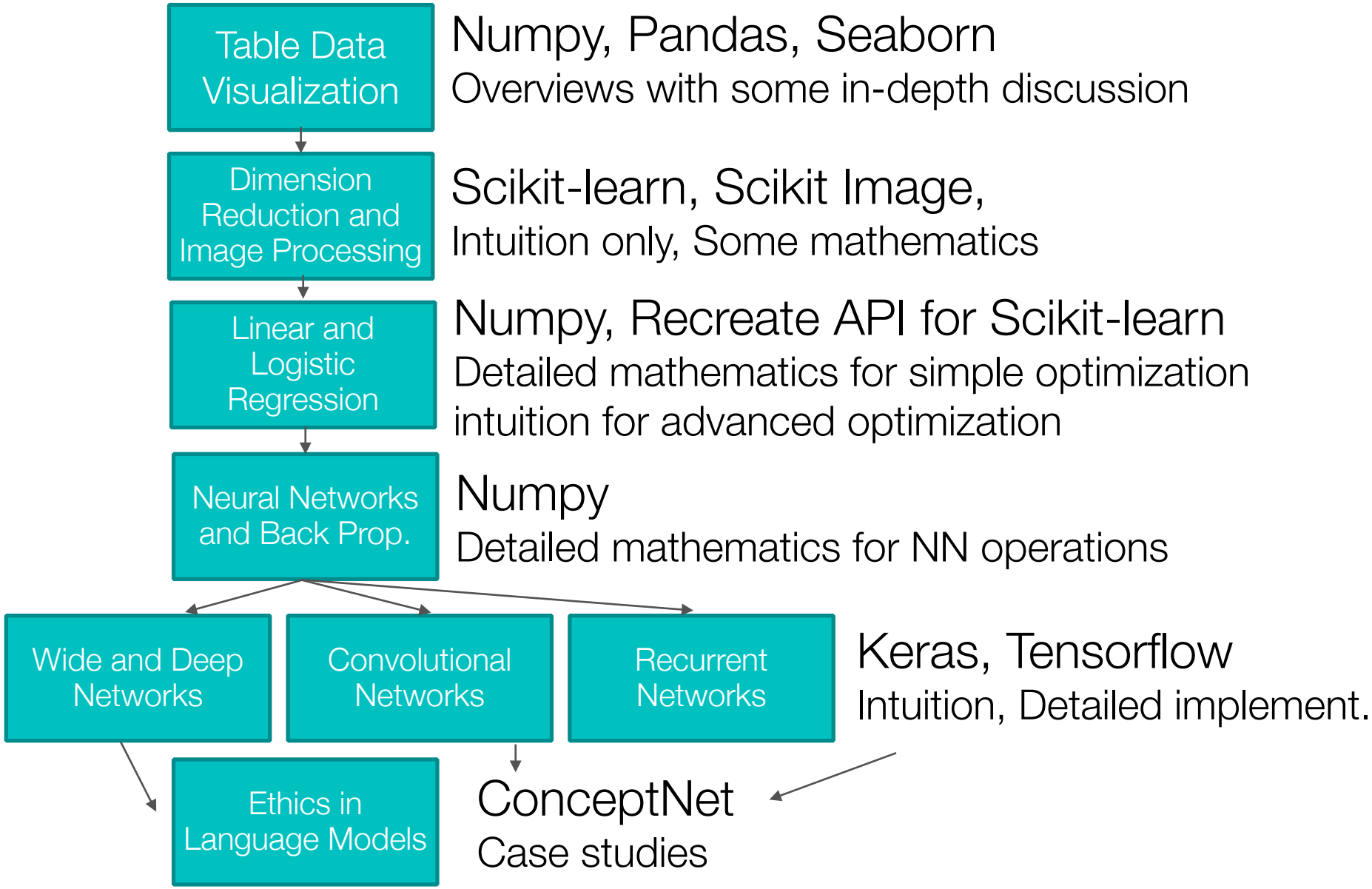
More graph interpretation



What's missing?

https://www.medrxiv.org/content/10.1101/2021.08.19.21262111v1.full.pdf

Class Overview, by topic



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

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!

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**

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

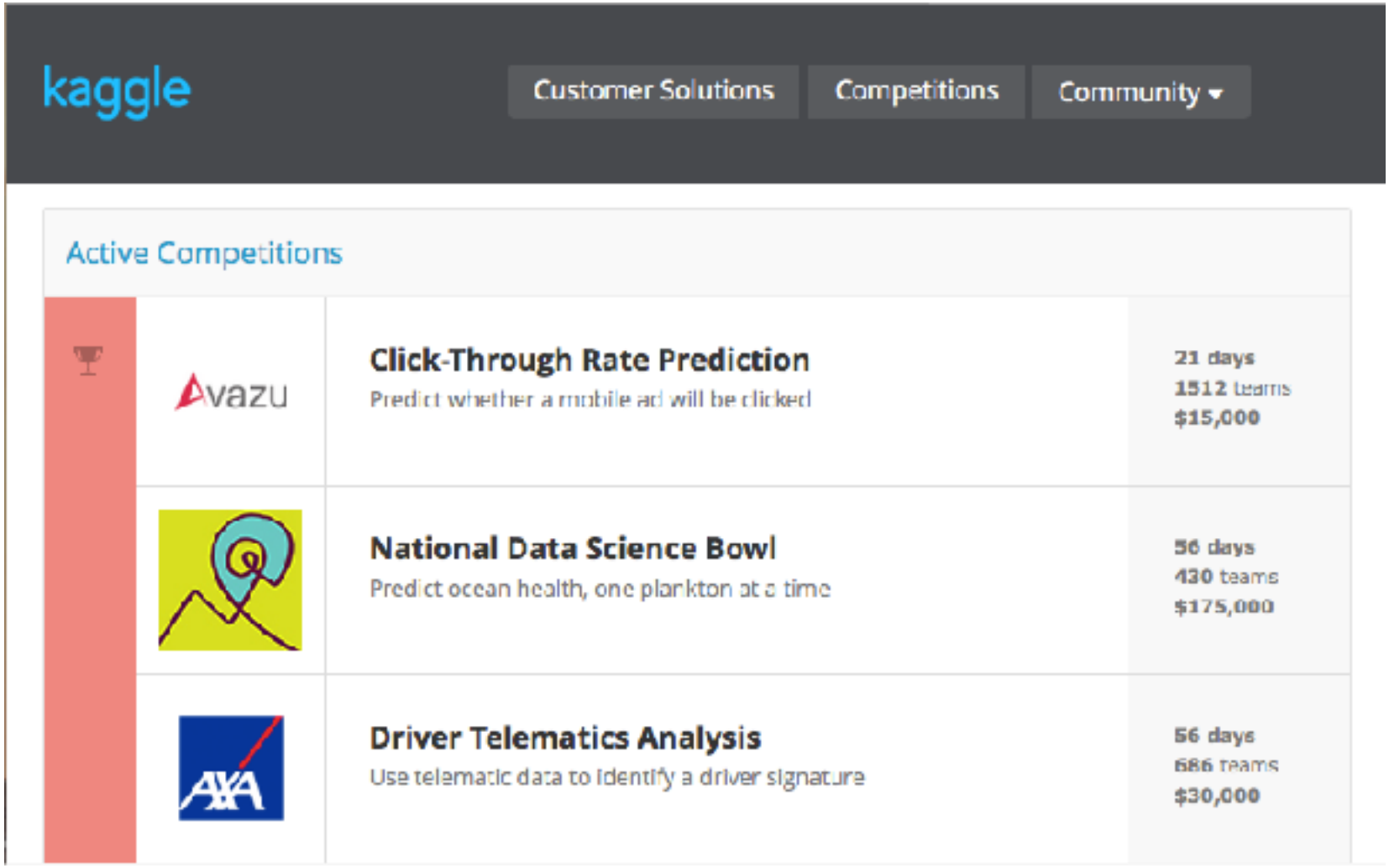
- Prediction Methods
  - Use some variables to predict unknown or future values of other variables
- Description Methods
  - Find human-interpretable patterns that describe the data.

ML

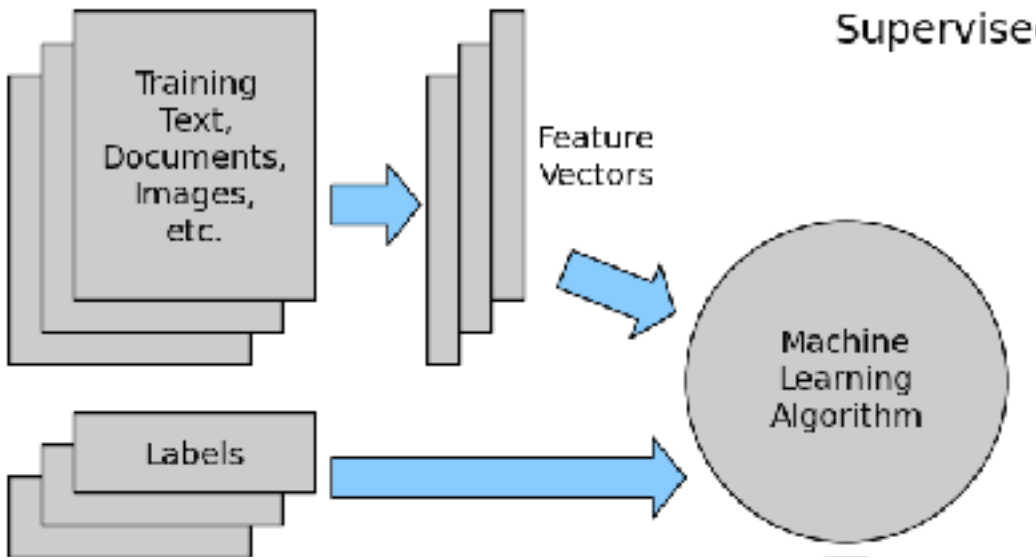
- Classification
- Regression
- Deviation Detection
- Clustering
- Association Rule Discovery
- Sequential Pattern Discovery

section 1, manipulated from Tan et al. Introduction to Data Mining

Problem Types in Machine Learning



Classification and Regression



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

TID	***Name***	***Size***
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

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

Unknown

TI	*# Rms*	*Sq Ft*	*Zip*
1	*2*	*2200*	*75115*