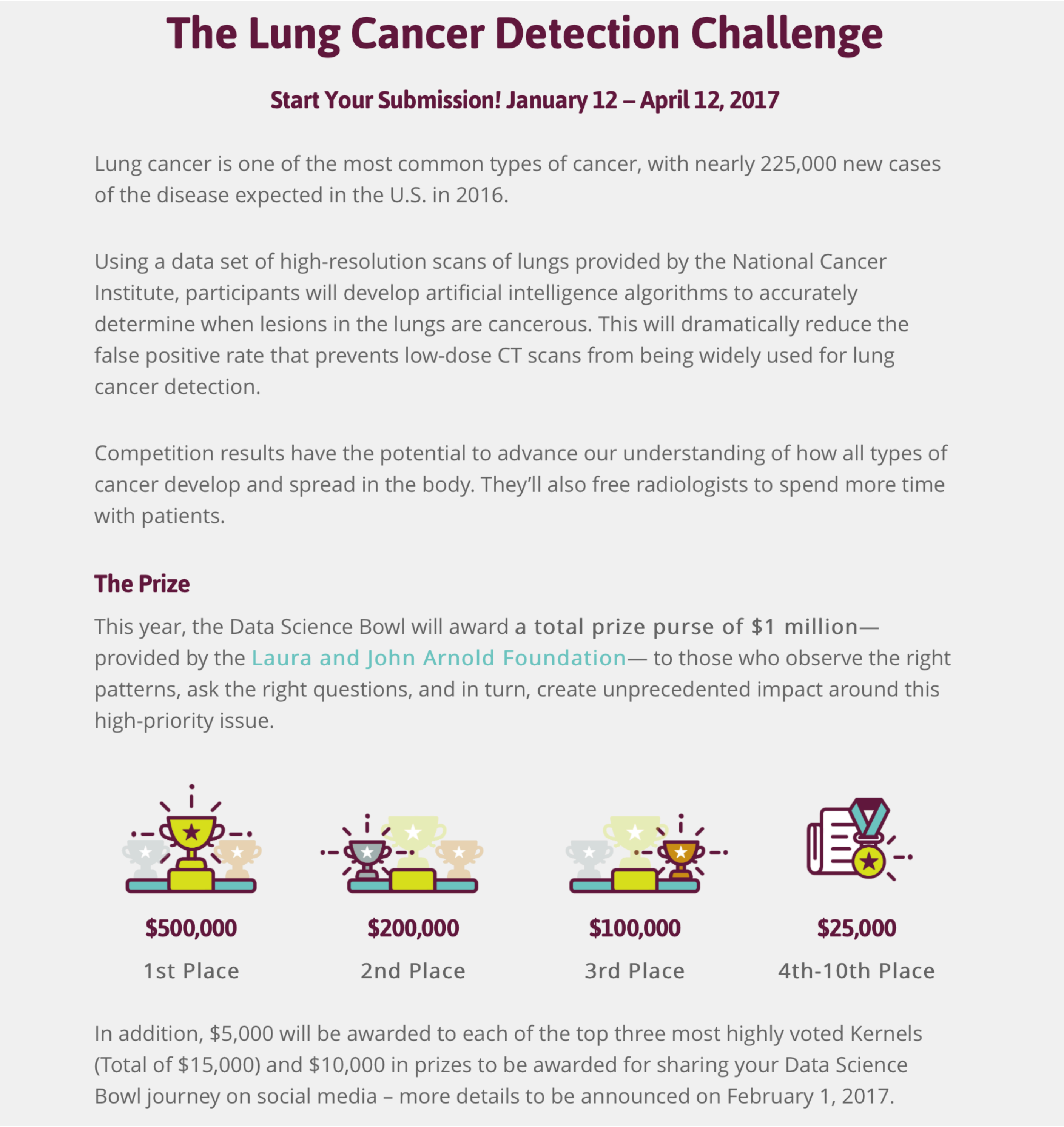
Machine Learning and Pattern Recognition

ECE 4950

Last month …

- BoozAllenHamilton & Kaggle announced
- http://www.datasciencebowl.com



The Lung Cancer Detection Challenge

Start Your Submission! January 12 – April 12, 2017

Lung cancer is one of the most common types of cancer, with nearly 225,000 new cases of the disease expected in the U.S. in 2016.

Using a data set of high-resolution scans of lungs provided by the National Cancer Institute, participants will develop artificial intelligence algorithms to accurately determine when lesions in the lungs are cancerous. This will dramatically reduce the false positive rate that prevents low-dose CT scans from being widely used for lung cancer detection.

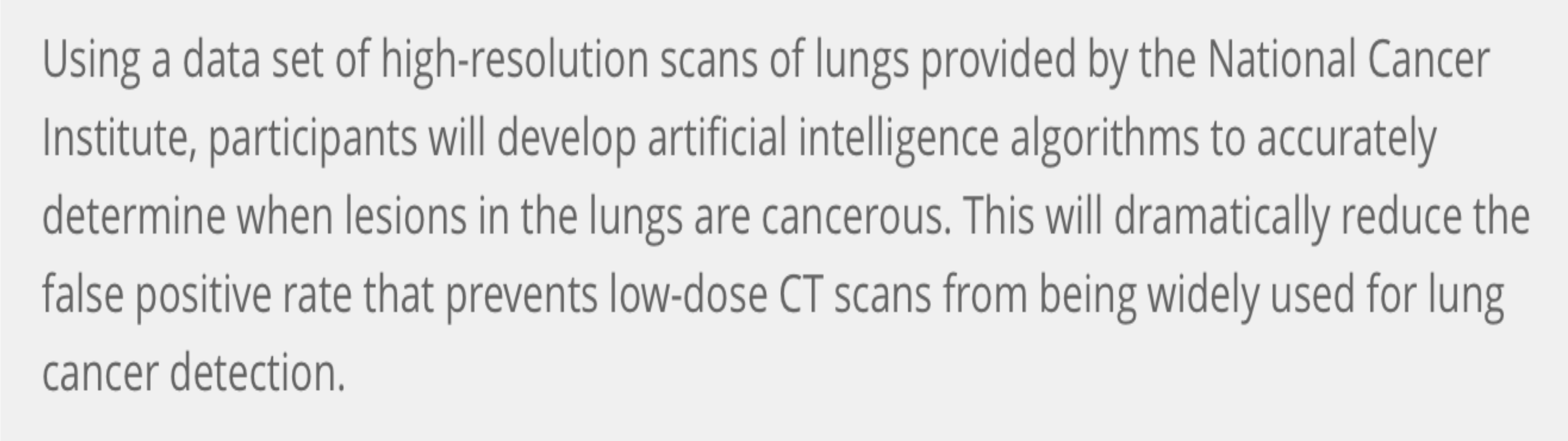
Competition results have the potential to advance our understanding of how all types of cancer develop and spread in the body. They'll also free radiologists to spend more time with patients.

The Prize

This year, the Data Science Bowl will award **a total prize purse of $1 million**— provided by the Laura and John Arnold Foundation— to those who observe the right patterns, ask the right questions, and in turn, create unprecedented impact around this high-priority issue.

$500,000	**$200,000**	**$100,000**	**$25,000**
1st Place	2nd Place	3rd Place	4th-10th Place

In addition, $5,000 will be awarded to each of the top three most highly voted Kernels (Total of $15,000) and $10,000 in prizes to be awarded for sharing your Data Science Bowl journey on social media – more details to be announced on February 1, 2017.



Using a data set of high-resolution scans of lungs provided by the National Cancer Institute, participants will develop artificial intelligence algorithms to accurately determine when lesions in the lungs are cancerous. This will dramatically reduce the false positive rate that prevents low-dose CT scans from being widely used for lung cancer detection.

Prof. Mert Sabuncu, ECE&BME

Team of Cornell students excited to participate

Prof. Sabuncu willing to be faculty mentor

http://blogs.cornell.edu/teams/2016-teams/datascience/

Please contact him if interested!

Will come back to this in a bit …

Hope: What we learn in this class will be helpful there …

Logistics

Lectures: MWF 11.15-12.05, Snee Hall Geological Sci 1150

Try as much **on board** as possible

Discussion: Tu 9.05-9.55 hours, Phillips Hall 407

Reinforce pre-requisites

Coding

Discussions are mandatory!

Logistics

Instructor: Jayadev Acharya, 304 Rhodes Hall

Office hours: MoTh 3-4 PM, Rhodes Hall 312

Teaching Assistant: Nirmal Vijay Shende

Office hours: TBA

https://people.ece.cornell.edu/acharya/teaching/ece4950s17/ece4950.html

WAITLIST on the website. Please put your names.

Prerequisites

Linear Algebra

Math2940 or equivalent

Basic Probability and Statistics

ECE3100, STSCI3080, ECE3250 or equivalent

Discussion session for reinforcing concepts
Not for introducing them!
Basic experience with python helpful!

Grading

- Assignments: 50%
  - 6-7 assignments, 1-2 weeks for turning in
  - Submission via CMS
- Miniproject: 25%
  - In-class Kaggle competition
  - Report and performance both matter
- Examination: 25%
  - Final examination

Will grade with other weights, and give better of the two grades

Websites

Class website linked from my website:

Piazza used for:

discussions, announcements, posting materials

www.piazza.com/cornell/spring2017/ece4950

CMS for turning in assignments

⚠

First time teaching such a class

Learn together

Please bear with inconveniences

MLPR – What are the problems?

Given examples (**training**), do:

- Decide something about new examples (**test**)

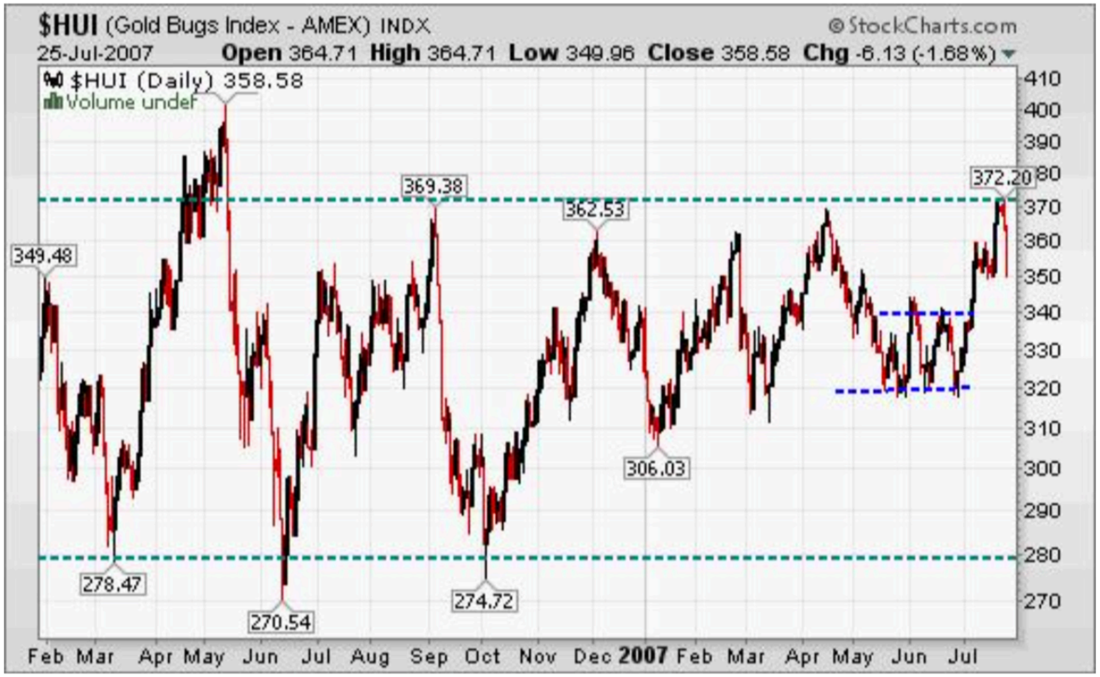
- Find interesting patterns in data

Example - Classification

e-mail-1: Spam
e-mail-2: Spam
e-mail-3: Ham
e-mail-4: Spam
…
e-mail-n: Ham

e-mailX = {Spam, Ham}?

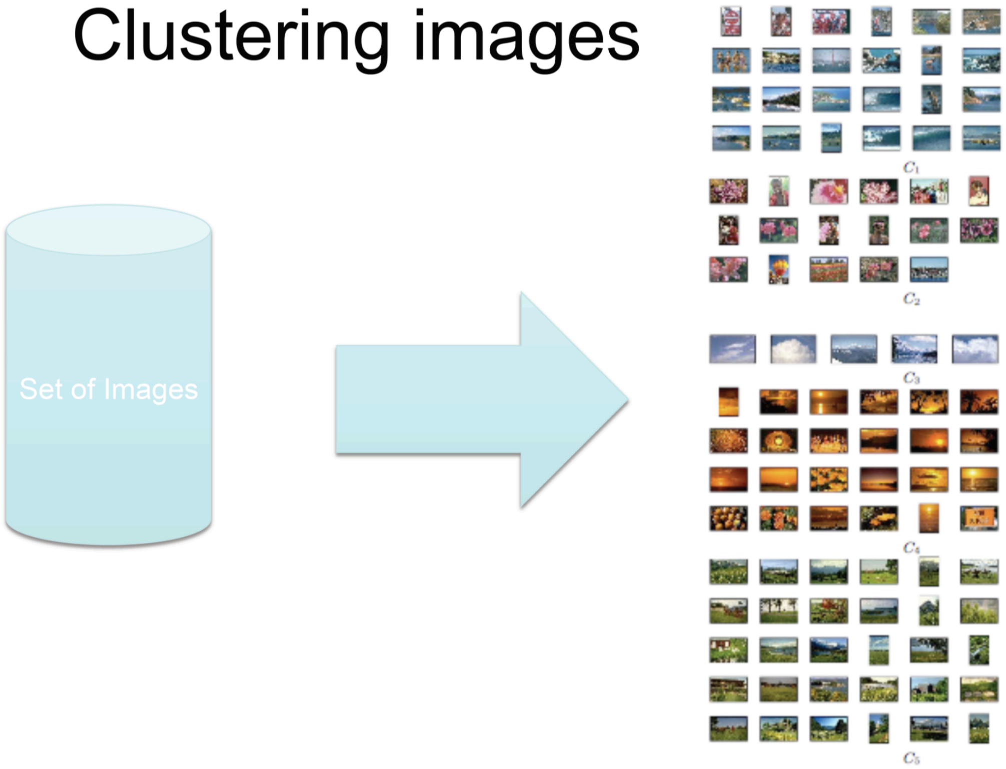
Example - Regression



Given past 1.5 years of Gold prices, predict tomorrow’s

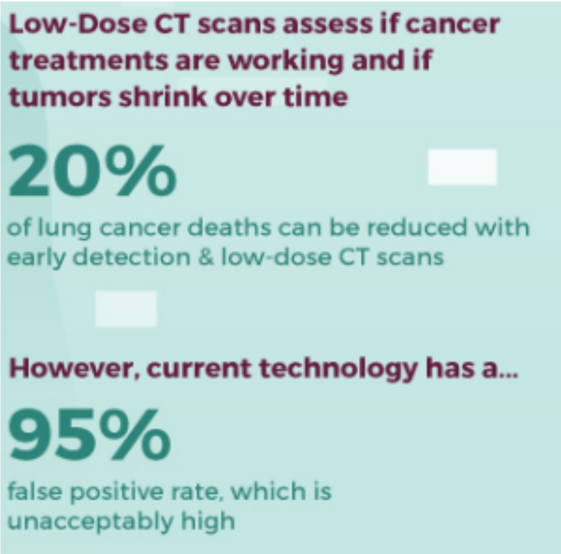
Example – Clustering

Clustering images



[Goldberger et al.]

Back to Data Science Bowl



Low-Dose CT scans assess if cancer treatments are working and if tumors shrink over time

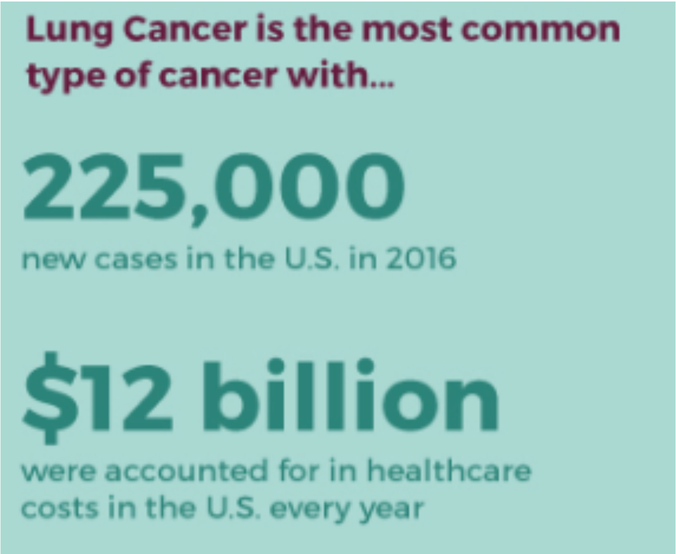
20%

of lung cancer deaths can be reduced with early detection & low-dose CT scans

However, current technology has a...

95%

false positive rate, which is unacceptably high



Lung Cancer is the most common type of cancer with...

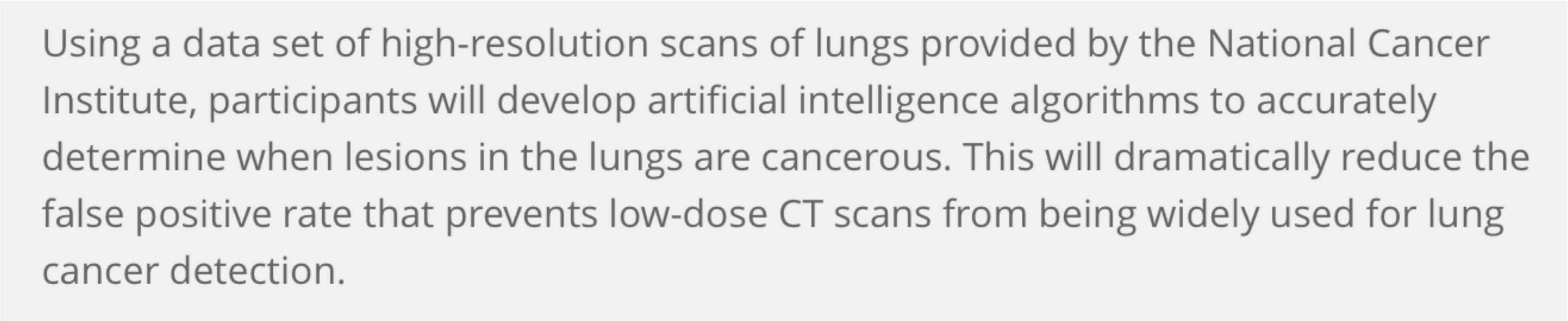
225,000

new cases in the U.S. in 2016

$12 billion

were accounted for in healthcare costs in the U.S. every year

Back to Data Science Bowl



Using a data set of high-resolution scans of lungs provided by the National Cancer Institute, participants will develop artificial intelligence algorithms to accurately determine when lesions in the lungs are cancerous. This will dramatically reduce the false positive rate that prevents low-dose CT scans from being widely used for lung cancer detection.

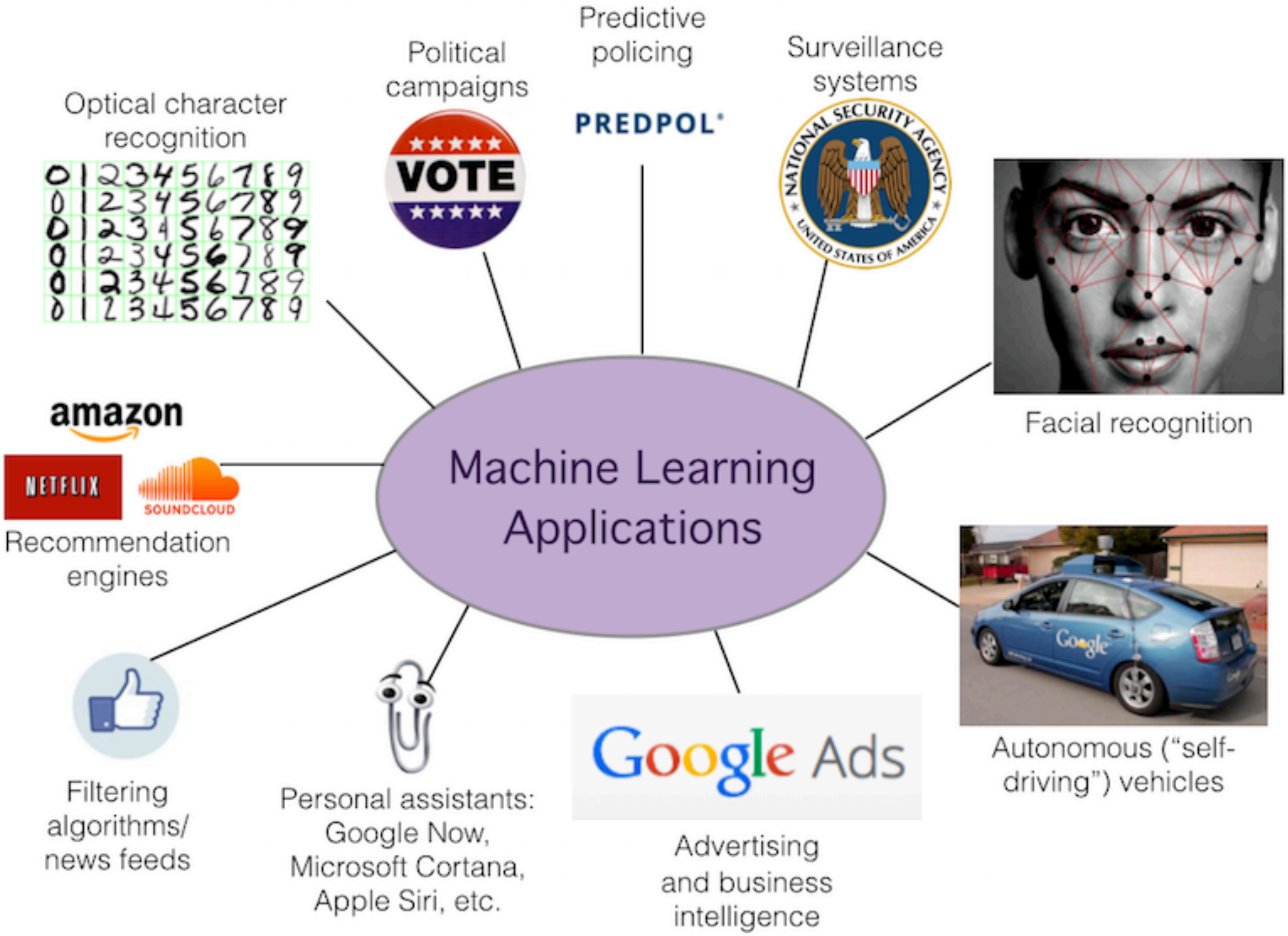
What do we have here?

Training: High resolution lung scans with labels (cancerous or not)

Test: Given new images, decide cancerous or not

GOOD LUCK!

Why learn this?


Optical character recognition
Political campaigns
VOTE
Predictive policing
PREDPOL
Surveillance systems
NATIONAL SECURITY AGENCY
UNITED STATES OF AMERICA
Facial recognition
amazon
NETFLIX
SOUNDCLOUD
Recommendation engines
Machine Learning Applications
Autonomous ("self-driving") vehicles
Google
Filtering algorithms/ news feeds
Personal assistants: Google Now, Microsoft Cortana, Apple Siri, etc.
Google Ads
Advertising and business intelligence

https://redshiftzero.github.io/assets/manip/ml_applications.png

MOVE TO THE BOARD now …