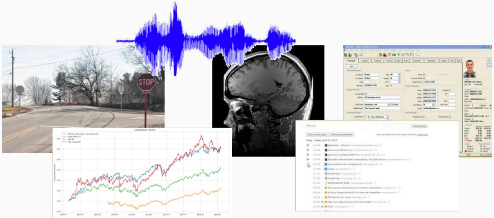
CS-UY 4563: Lecture 1
Introduction to Machine Learning

NYU Tandon School of Engineering, Prof. Christopher Musco

BASIC GOAL

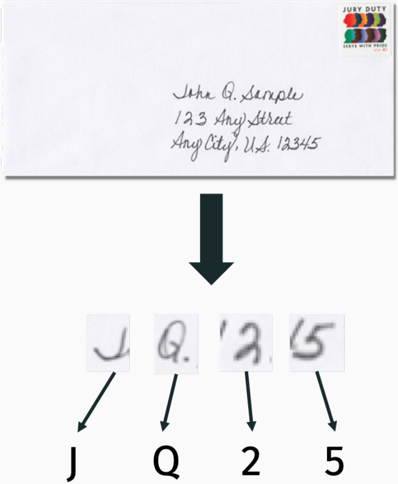
Goal: Develop algorithms to make decisions or predictions based on data.

- **Input:** A single piece of data (an image, audio file, patient healthcare record, MRI scan).



- **Output:** A prediction or decision (this image is a stop sign, this stock will go up 10% next quarter, turn the car right).

Optical character recognition (OCR): Decide if a handwritten character is an $a, b, \ldots, z, 0, 1, \ldots, 9, \ldots$.



Optical character recognition (OCR): Decide if a handwritten character is an $a, b, \ldots, z, 0, 1, \ldots, 9, \ldots$.

Applications:

- Automatic mail sorting.
- Text search in handwritten documents.
- Digitizing scanned books.
- License plate detection for tolls.
- Etc.

How would you write an **algorithm** to distinguish these digits?

Suppose you just want to distinguish between a 1 and a 7.

Reasonable approach: A number which contains one vertical line is a 1, if it contains one vertical and one horizontal line, it's a 7.

```python
def count_vert_lines(image):
...

def count_horiz_lines(image):
...

def classify(image):
...
    nv = count_vert_lines(image)
    nh = count_vert_lines(image)

    if (nv == 1) and (nh == 1):
        return '7'
    elif (nv == 1) and (nh == 0):
        return '1'
    elif ...
```

This rule breaks down in practice:

Even fixes/modifications of the rule tend to be brittle... Maybe you could get 80% accuracy, but not nearly good enough.

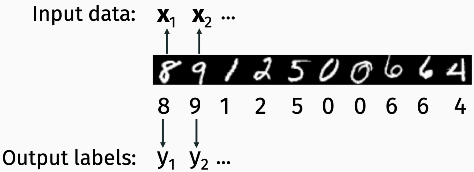
Rule based systems, also called Expert Systems were the dominant approach to artificial intelligence in the 70s and 80s.

Major limitation: While human's are very good at many tasks,

- It's often hard to encode why humans make decisions in simple programmable logic.
- We think in abstract concepts with no mathematical definitions (how exactly do you define a line? how do you define a curve? straight line?)

Focus on what humans do well: solving the task at hand!

Step 1: Collect and label many input/output pairs $(\mathbf{x}_i, y_i)$. For our digit images, we have each $\mathbf{x}_i \in \mathbb{R}^{28\times 28}$ and $y_i \in \{0, 1, \ldots, 9\}$.



This is called the training dataset.

Step 2: Learn from the examples we have.

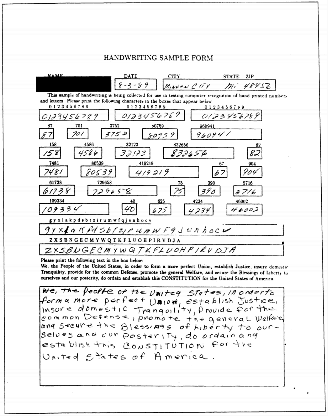
- Have the computer automatically find some function $f(\mathbf{x})$ such that $f(\mathbf{x}_i) = y_i$ for most $(\mathbf{x}_i, y_i)$ in our training data set (by searching over many possible functions).

Think of f as any crazy equation, or an arbitrary program:

$$f(\mathbf{x}) = 10 \cdot \mathbf{x}[1,1] - 6 \cdot \mathbf{x}[3,45] \cdot \mathbf{x}[9,99] + 5 \cdot \text{mean}(\mathbf{x}) + \ldots$$

This approach of learning a function from labeled data is called **supervised learning.**

National Institute for Standards and Technology collected a huge amount of handwritten digit data from census workers and high school students in the early 90s:



HANDWRITING SAMPLE FORM

NAME [redacted] DATE 8-3-89 CITY MINDEN CITY STATE MI ZIP 48456

This sample of handwriting is being collected for use in testing computer recognition of hand printed numbers and letters. Please print the following characters in the boxes that appear below

0 1 2 3 4 5 6 7 8 9   0 1 2 3 4 5 6 7 8 9   0 1 2 3 4 5 6 7 8 9

87   701   3752   80759   960941

158   4586   32123   832656   82

7481   80539   419219   67   904

61738   729658   75   390   5716

109334   40   625   4234   46002

g y x l a k p d s b t z i r u m w f q j e n h o c v

Z X S B N G E C M Y W Q T K F L U O H P I R V D J A

Please print the following text in the box below:
We, the People of the United States, in order to form a more perfect Union, establish Justice, insure domestic Tranquility, provide for the common Defense, promote the general Welfare, and secure the Blessings of Liberty to ourselves and our posterity, do ordain and establish this CONSTITUTION for the United States of America

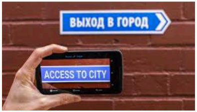
This is called the NIST dataset, and was used to create the famous MNIST handwritten digit dataset.

Since the 1990s machine learning have overtaken expert systems as the dominant approach to artificial intelligence.

- Current methods achieve .21% error rate for OCR on benchmark datasets (MNIST).
- Very successful on other problems as well.

You could not be studying ML at a more exciting time!

- Autonomous vehicles.
- Human level play in very difficult games.
- Incredible machine translation.
- Pervasive impact in science and engineering.
- Many, many more.



Machine learning has benefited from an explosion in our ability to collect and store data:

- Cheap, fast storage. Large data centers.
- Pervasive monitoring (satellite imagery, cheap sensors, improved and reduced cost for technologies like LIDAR).
- Crowd-sourced data collection (images, text on the internet)
- Crowd-sourced data labeling via the internet (Amazon Mechanical Turk, reCAPTCHA, etc.)

Having lots of data isn't enough. We have to know how to use it effectively.

CENTRAL QUESTIONS IN MACHINE LEARNING

Once we have the basic machine learning setup, many very difficult questions remain:

- How do we parameterize a class of functions f to search?
- How do we efficiently find a good function in the class?
- How do we ensure that an $f(\mathbf{x})$ which works well on our training data will generalize to perform well on future data?
- How do we deal with imperfect data (noise, outliers, incorrect training labels)?

In this course you will learn to answer these central questions through a combination of:

- Hands on implementation.
  - In-class demos and take-home labs using `Python` and `Jupyter notebooks`.
  - Final Project (on any dataset/problem you like).
- Theoretical exploration.
  - Written problem sets.
  - Two midterm exams.

Goals of hands-on component:

1. Learn how to view and formulate real world problems in the language of machine learning.
2. Gain experience applying the most popular and most successful machine learning algorithms to example problems. The goal is to prepare you to use these tools in industrial or academic positions.

Goals of theoretical component:

1. Learn how theoretical analysis can help explain the performance of machine learning algorithms and lead to the design of entirely new methods.
2. Build experience with the most important mathematical tools used in machine learning, including probability, statistics, and linear algebra. This experience will prepare you for more advanced coursework in ML, or research.
3. Be able to understand contemporary research in machine learning, including papers from NeurIPS, ICML, ICLR, and other major machine learning venues.

All class information can be found at:

`www.chrismusco.com/introml`