# CSC 589: Introduction to Machine Learning

**Instructor:** Dr. Nathalie Japkowicz

**e-mail:** japkowicz@american.edu



## Objectives of the Course And Preliminaries

# Some Information

- **Instructor:** Dr. Nathalie Japkowicz
- **Office:** DMTI-112B
- **Phone Number:** (202) 885-6486 (don't rely on it too much!)
- **E-mail:** japkowic@american.edu (best way to contact me!)
- **Office Hours:**
  - Thursdays, 2:30pm-4pm
  - By arrangement, on Skype

# Machine Learning: A Case Study

- Malfunctioning gearboxes have been the cause for CH-46 US Navy helicopters to crash.
- Although gearbox malfunctions can be diagnosed by a mechanic prior to a helicopter's take off, what if a malfunction occurs while in-flight, when it is impossible for a human to detect?



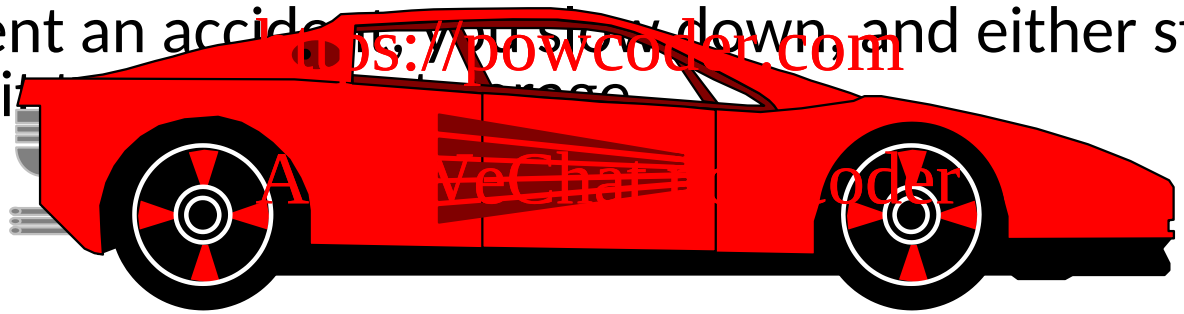
- Machine Learning was shown to be useful in this domain and thus to have the potential of saving human lives!

# How did it Work?



Consider the following common situation:

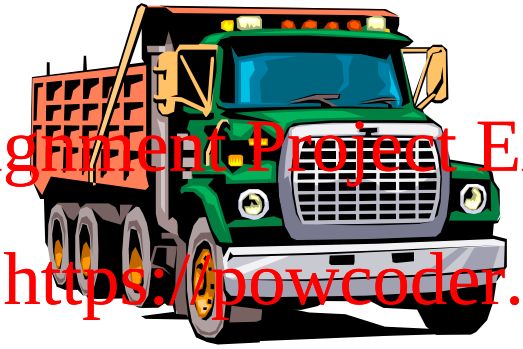
- You are in your car, speeding away, when you suddenly hear a "funny" noise.
- To prevent an accident you slow down, and either stop the car or bring it



- The in-flight helicopter gearbox fault monitoring system was designed following the same idea. The difference, however, is that many gearbox malfunction cannot be heard by humans

# So, Where's the Learning?

- Imagine that, instead of driving your good old battered car, you were asked to drive this truck:



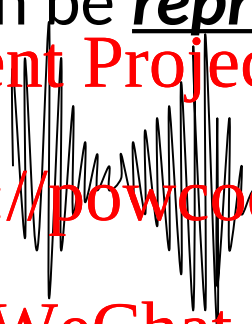


- Would you know a "funny" noise from a "normal" one?
- Well, probably not, since you've never driven a truck before!
- While you drove your car during all these years, you effectively ***learned*** what your car sounds like and this is why you were able to identify that "funny" noise.

# What did the Computer Learn?

- Obviously, a computer cannot hear and can certainly not distinguish between a normal and an abnormal sound.
- Sounds, however, can be ***represented*** as wave patterns such as this one:





which in fact is a series of real numbers indicating intensity.

- And computers can deal with strings of numbers!
- For example, a computer can easily be programmed to distinguish between strings of numbers that contain a “3” in them and those that don’t.

# What did the Computer Learn? (Cont'd)

- In the helicopter gearbox monitoring problem, the assumption is that functioning and malfunctioning gearboxes emit different sounds. Thus, the strings of numbers that represent these sounds have different characteristics.
- The exact characteristics of these different categories, however, are unknown and/or are too difficult to describe.
- Therefore, they cannot be programmed, but rather, they need to be learned by the computer.
- There are many ways in which a computer can learn how to distinguish between two patterns (e.g., decision trees, neural networks, bayesian networks, etc.) and that is the topic of this course!

# What else can Machine Learning do?

- Medical Diagnostic (e.g., breast cancer detection)
- Credit Card Fraud Detection
- Sonar Detection (e.g., submarines versus shrimps (!) )
- Speech Recognition (e.g., Telephone automated systems)
- Autonomous Vehicles (useful for hazardous missions or to assist disabled people)
- Personalized Web Assistants (e.g., an automated assistant can assemble personally customized newspaper articles)
- And many more applications…

# Text Books and Reading Material

- **Peter Flach**, *Machine Learning: The art and science of algorithms that make sense of data*. Cambridge University Press, 2012.
- **Nathalie Japkowicz and Mohak Shah**, *Evaluating Learning Algorithms: A Classification Perspective* , Cambridge University Press, 2011.
- **Research papers** saved in the directory entitled literature on Blackboard
- On Blackboard, you will also find a list of non-required books that you may find useful.

# Objectives of the Course:

- To present a broad introduction of the principles and paradigms underlying machine learning, including discussions and hands-on evaluations of some of the major approaches currently being investigated.
- To introduce the students to the reading, presenting and critiquing of research papers.
- To initiate the students to formulating a research problem and carrying this research through.

# Format of the Course:

- The course is lecture based.
- Each student will write 6 research paper critiques as part of a group of 2 or 3 students.
- Each student is expected to present, in class, one of the 13 papers provided in the literature packet. (Each student will present a different paper). The student presentations will take place during the weeks of **November 26** and **December 3**.
- On the **last day of classes**, each student will present a poster to the entire class (and the whole department). This poster will be based on the research they will have carried out for their final project.
- There will be an in-class midterm exam on **November 1**, but no final exam.
- There will be two assignments and a final project.

# Contents of the Course I :

- The course will teach machine learning algorithms, theoretical issues and contemporary problems in machine learning.
- Machine learning algorithms covered:
  - Version Spaces
  - Decision Trees
  - Artificial Neural Networks
  - Bayesian Learning
  - Instance-Based Learning
  - Support Vector Machines
  - Ensemble-Learning Algorithms
  - Rule Learning/Associative Rule Mining
  - Unsupervised Learning/Clustering

# Contents of the Course II:

- Theoretical issues considered:
  - The roots of Machine learning (Philosophy, AI, Computational Learning Theory, Statistics)
  - Empirical issues: Evaluation of Learning Algorithms
- Practical issues considered:
  - Data Exploration
  - Data Preparation
  - Feature Selection
  - The Class Imbalance Problem

# Contents of the Course III:

Contemporary themes considered:

We will be specifically looking at the area of big data analysis, and, possibly, at some of the following topics:



- Deep Learning
- Graph Mining



- Mining Social Networks
- Data Streams Mining



- Unstructured or Semi-Structured Data Mining
- Data Mining with Heterogeneous Sources
- Spatio-Temporal Data Mining
- Issues of Trust and Provenance in Data Mining
- Privacy in Data Mining
- Cybersecurity

# Course Requirements:

**Percent of the Final Grade**

- Written commentaries — 12%
- Oral presentation of a research paper — 8%
- 2 Assignments (little programming involved as programming packages will be provided) — 20%
- Midterm Exam — 25%
- Final Project: - Project Proposal
  - Project Report
  - Poster Presentation — 35%

# More on Assignments and Final Exam:



- **Assignment 1:**
  - Handed out on: Thursday September 27, 2018
  - Due on: Thursday Oct 18, 2018
- **Assignment 2:**
  - Handed out on: Thursday Oct 18, 2018
  - Due on: Thursday November 15, 2018

- **Midterm Exam:**
  - In class on Thursday November 1, 2018

# More on the Research Project (Also see the Project Description on Blackboard)

- **Research Project** including a literature review and the design and implementation of a novel learning scheme or the comparison of several existing schemes.
- **Projects Proposal** (3-5 pages) are due on **October 25, 2018**
- **Project Report** are due on **December 6, 2018**
- **Project Presentations** will take place on **December 6, 2018** in the form of a poster presentation.
- Suggestions for project topics are listed on Blackboard, but you are welcome (and that's even better) to propose your own idea.

**Start thinking about the project early!!!!!**