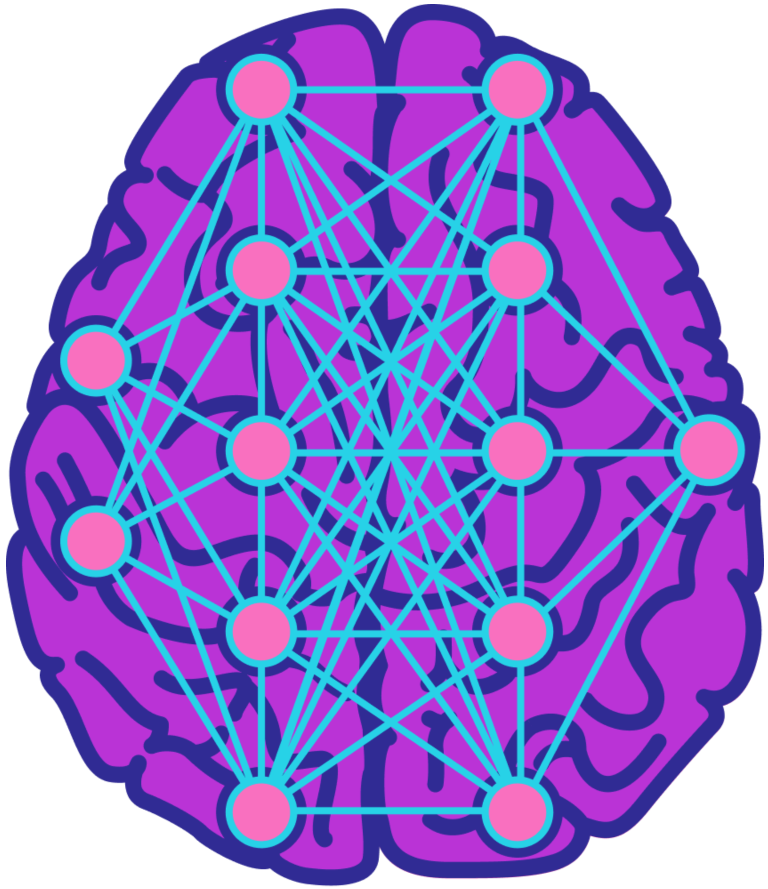
DS 4400

Machine Learning and Data Mining I

Alina Oprea
Associate Professor, CCIS
Northeastern University

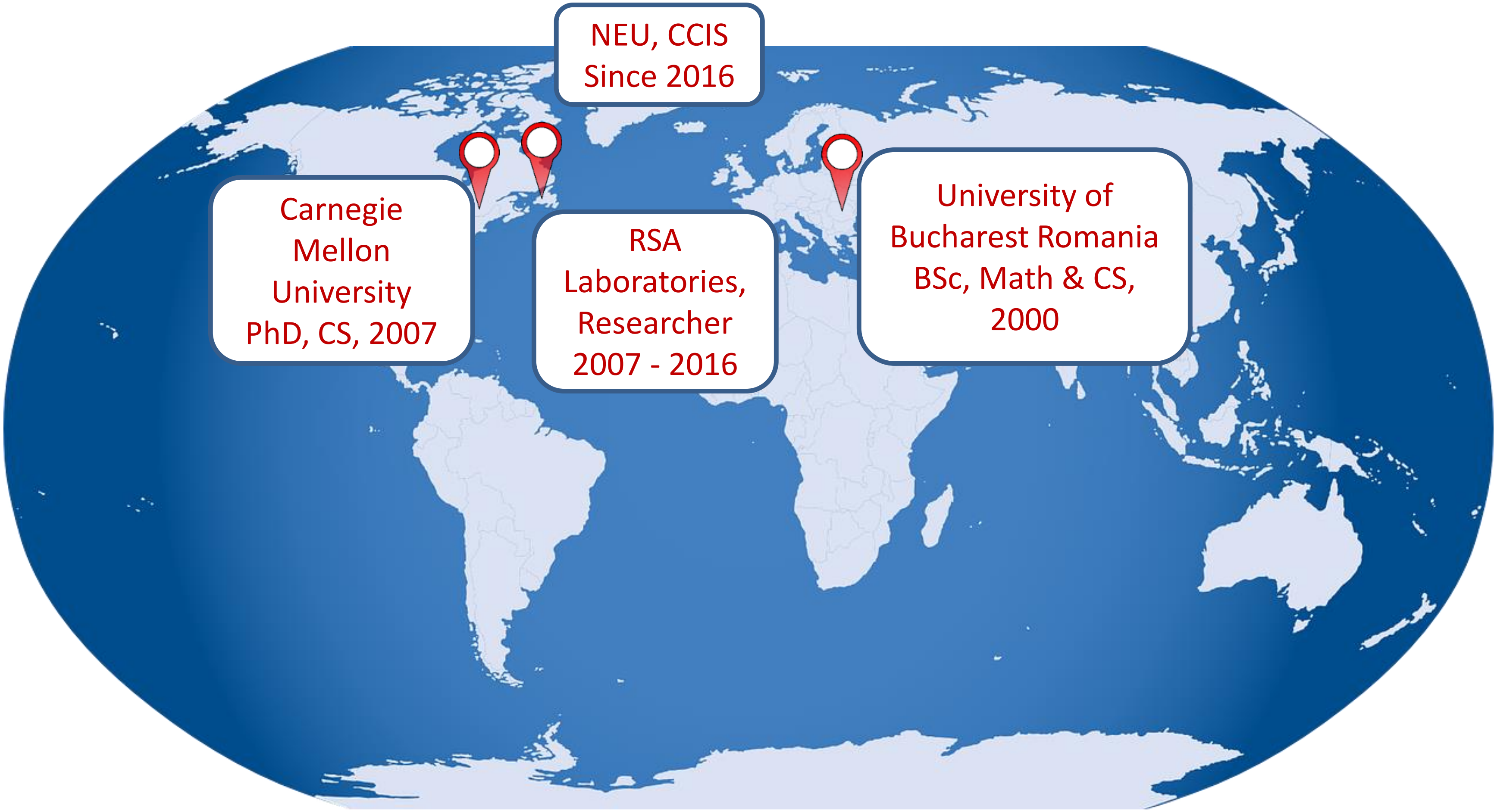
January 8 2019

Welcome to DS 4400!



Machine Learning and Data Mining I

Introductions



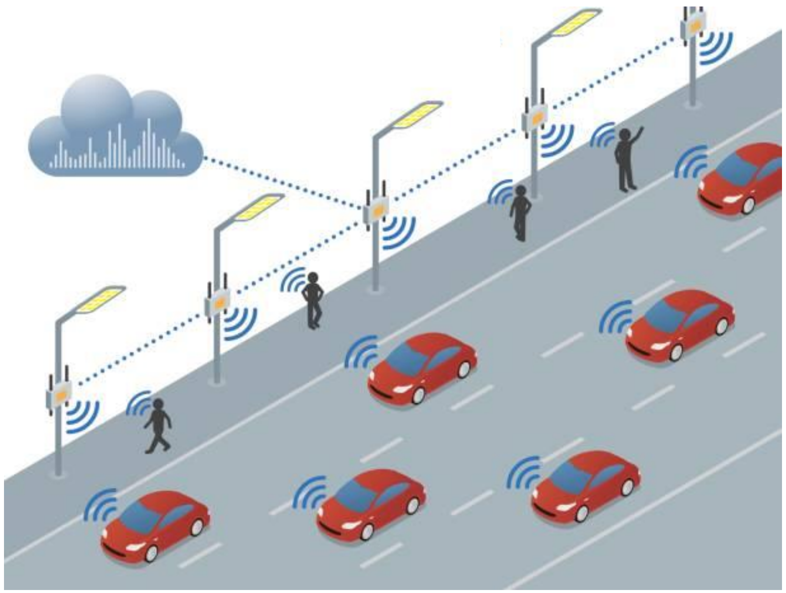
Background

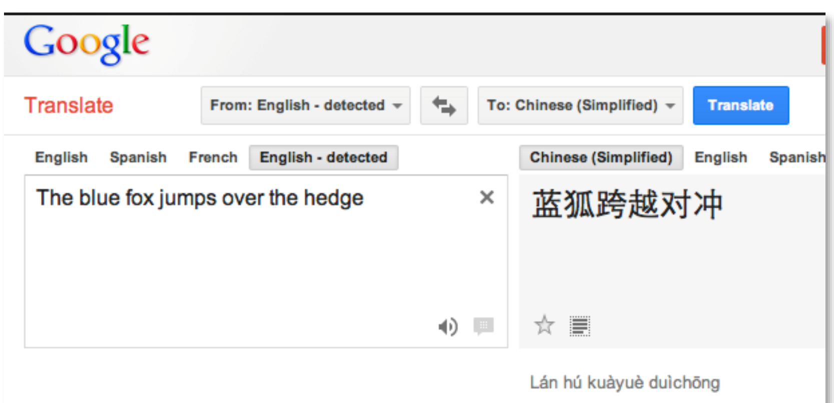
- Ph.D. at CMU
  - Research in storage security & cryptographic file systems
- RSA Laboratories
  - Cloud security, applied cryptography
  - Security analytics (ML in security)
- NEU CCIS – since Fall 2016
  - ML for security applications (attack detection, IoT, connected car security)
  - Adversarial ML (study the vulnerabilities of ML in face of attacks and design defenses)

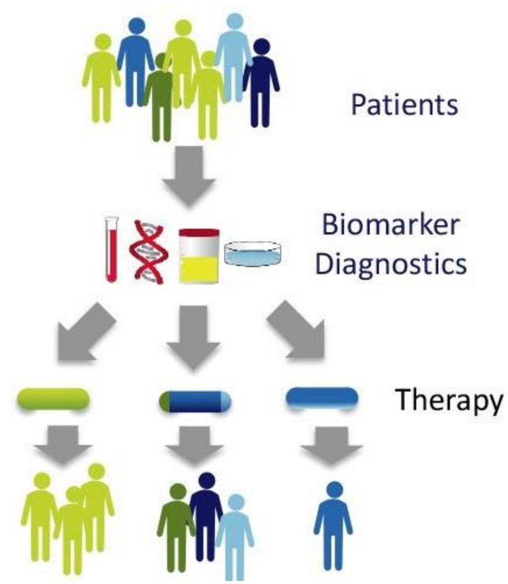
Class Introductions

- Enrollment of 27
- Diverse majors
  - CCIS
  - Math
  - Economics
  - Biology
  - Engineering

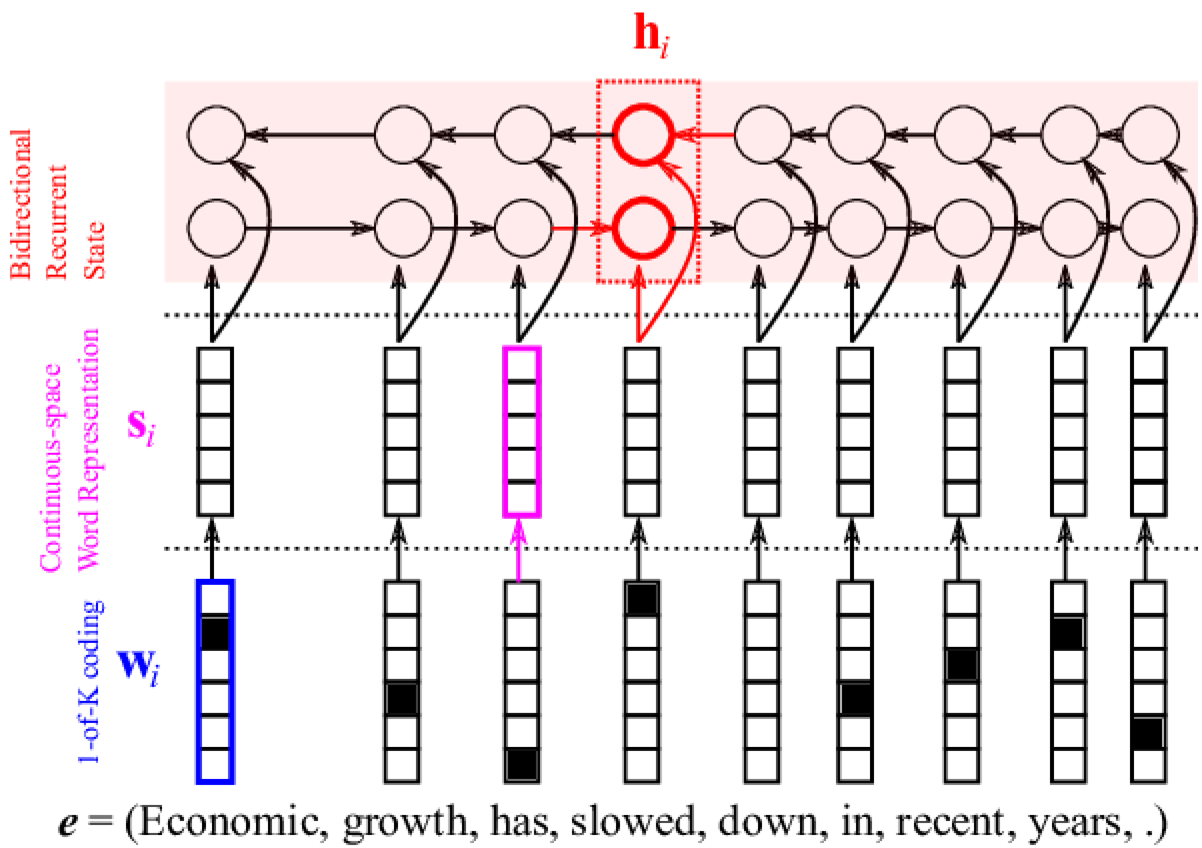
Machine learning is everywhere





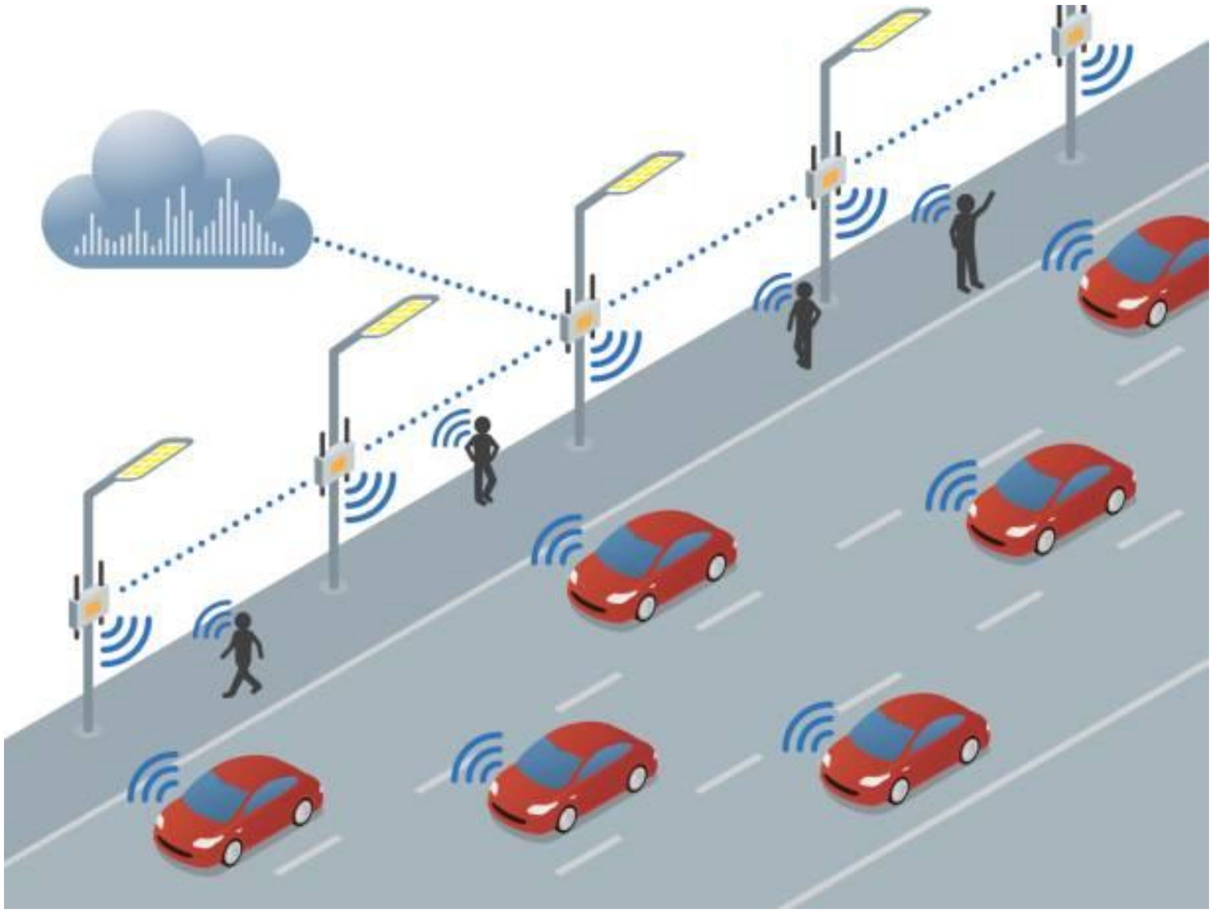


Natural Language Processing (NLP)



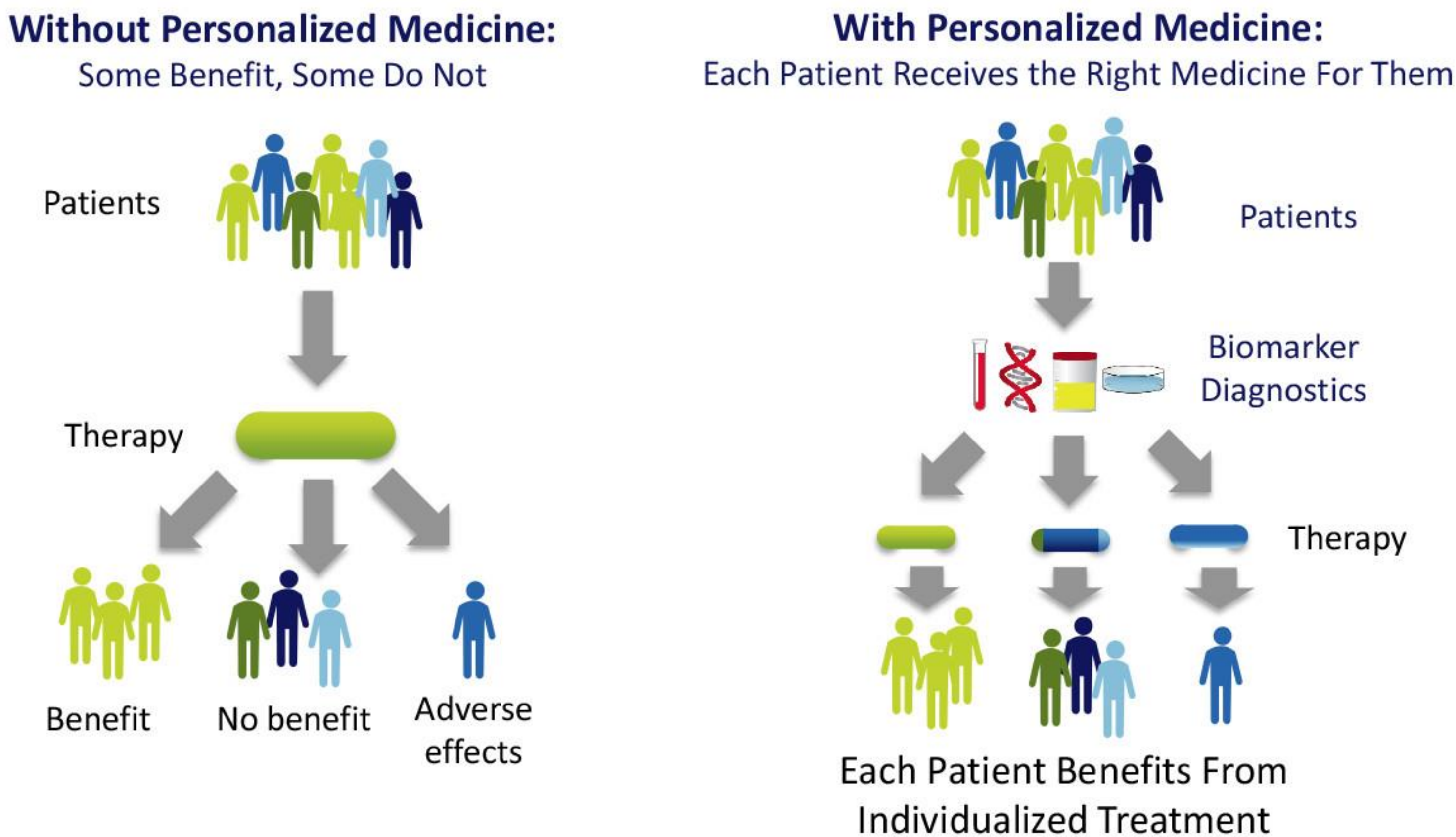
- Understand language semantics
- Real-time translation, speech recognition

Autonomous vehicles



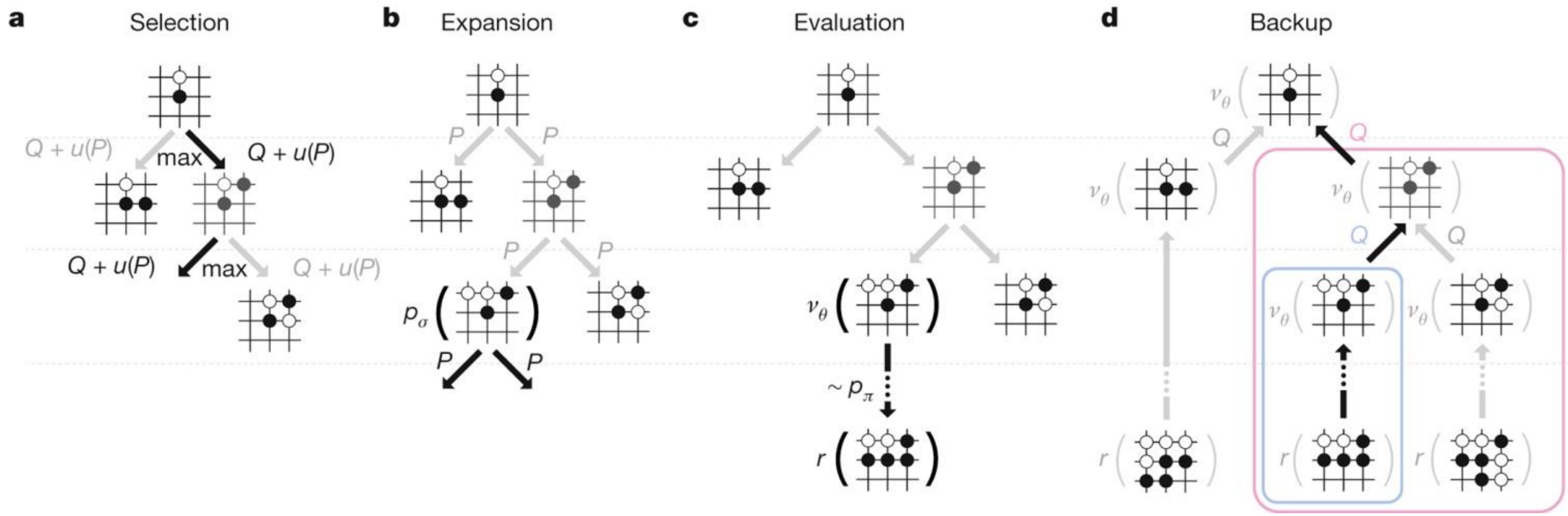
- Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) communication
- Assist drivers in making decisions to increase safety

Personalized medicine



- Treatment adjusted to individual patients
- Predictive models using a variety of features related to patient history and genetics

Playing games



Reinforcement learning

- AlphaGo
- Chess

DS-4400

- What is *machine learning*?
  - The science of teaching machines how to learn
  - Design predictive algorithms that learn from data
  - Replace humans in critical tasks
  - Subset of Artificial Intelligence (AI)
- Machine learning very successful in:
  - Machine translation
  - Precision medicine
  - Recommendation systems
  - Self-driving cars
- Why the hype?
  - Availability: data created/reproduced in 2010 reached 1,200 exabytes
  - Reduced cost of storage
  - Computational power (cloud, multi-core CPUs, GPUs)

DS-4400 Course objectives

- Become familiar with machine learning tasks
  - Supervised learning vs unsupervised learning
  - Classification vs Regression vs Clustering
- Study most well-known algorithms and understand their details
  - Regression (linear regression)
  - Classification (SVM, decision trees, neural networks)
  - Clustering (k-means, hierarchical clustering)
- Learn to apply ML algorithms to real datasets
  - Using existing packages in R and Python
- Learn about security challenges of ML
  - Introduction to adversarial ML

http://www.ccs.neu.edu/home/alina/classes/Spring2019

Class Outline

- Introduction – 1 week
  - Probability and linear algebra review
- Supervised learning - 7 weeks
  - Linear regression
  - Classification (logistic regression, LDA, kNN, decision trees, random forest, SVM, Naïve Bayes)
  - Model selection, regularization, cross validation
- Neural networks and deep learning – 2 weeks
  - Back-propagation, gradient descent
  - NN architectures (feed-forward, convolutional, recurrent)
- Unsupervised learning – 1-2 weeks
  - Dimensionality reduction (PCA)
  - Clustering (k-means, hierarchical)
- Adversarial ML – 1 lecture
  - Security of ML at testing and training time

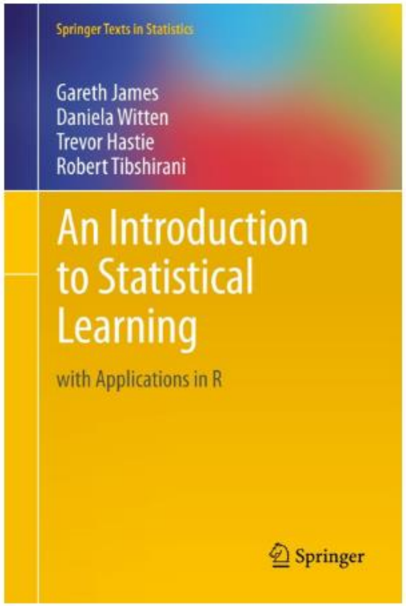
Textbook

An Introduction to Statistical Learning

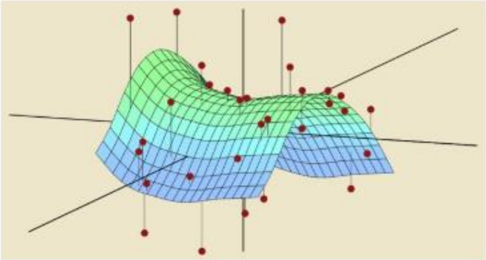
with Applications in R

Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani

- **Home**
- **About this Book**
- **R Code for Labs**
- **Data Sets and Figures**
- **ISLR Package**
- **Get the Book**
- **Author Bios**
- **Errata**



Download the book PDF
(corrected 7th printing)



Statistical Learning MOOC covering the entire ISL book offered by Trevor Hastie and Rob Tibshirani. Start anytime in self-paced mode.

This book provides an introduction to statistical learning methods. It is aimed for upper level undergraduate students

Specific chapters will be covered

Other resources

- Trevor Hastie, Rob Tibshirani, and Jerry Friedman, Elements of Statistical Learning, Second Edition, Springer, 2009.

- Christopher Bishop. Pattern Recognition and Machine Learning. Springer, 2006.

- Ian Goodfellow and Yoshua Bengio and Aaron Courville. Deep Learning. MIT Press. 2016

Policies

- Instructors
  - Alina Oprea
  - TA: Ewen Wang
- Schedule
  - Tue 11:45am – 1:25pm, Thu 2:50-4:30pm
  - Shillman Hall 210
  - Office hours:
    - Alina: Thu 4:30 – 6:00 pm (ISEC 625)
    - Ewen: Monday 5:30-6:30pm (ISEC 605)
- Online resources
  - Slides will be posted after each lecture
  - Use Piazza for questions, Gradescope for homework and project submission

Policies, cont.

- Your responsibilities
  - Please be on time, attend classes, and take notes
  - Participate in interactive discussion in class
  - Submit assignments/ programming projects on time
- Late days for assignments
  - 5 total late days, after that loose 20% for every late day
  - Assignments are due at 11:59pm on the specified date
  - No need to email for late days, Gradescope shows submission time

Grading

- Assignments – 25%
  – 4-5 assignments and programming exercises based on studied material in class
- Final project – 35%
  – Select your own project based on public dataset
  – Submit short project proposal and milestone
  – Presentation at end of class (10 min) and report
- Exam – 35%
  – One exam about 3/4 in the class
  – Tentative end of March
- Class participation – 5%
  – Participate in class discussion and on Piazza

Assignments

- Mostly programming exercises, occasionally some theory questions
- Language
  - Use R or Python
  - Jupyter notebooks recommended
- Submission
  - Submit PDF report in Gradescope
  - Includes all the results, as well as link to code and instructions to run it

Final project

- Goal: work on a larger data science project
  - Build your portfolio and increase your experience
- Requirements
  - Large dataset: at least 10,000 records (public source)
  - Not recommended to collect your own data
  - Pick application of interest, but instructor will also provide potential list of projects
  - Experiment with at least 3 ML models
  - Perform in-depth analysis (which features contribute mostly to prediction, which model performs best)
- Timeline
  - Proposal: mid class; milestone 2-3 weeks after (Instructor will provide early feedback)
  - Final presentation (10 mins) and report (5-6 pages)

Academic Integrity

- Homework is done individually!
- Final project is done individually!
- Rules
  - Can discuss with colleagues or instructor
  - Can post and answer questions on Piazza
  - Code cannot be shared with colleagues
  - Cannot use code from the Internet
    - Use python or R packages, but not directly code for ML analysis written by someone else
- NO CHEATHING WILL BE TOLERATED!
- Any cheating will automatically result in grade F and report to the university administration
- http://www.northeastern.edu/osccr/academic-integrity-policy/

Outline

- Supervised learning
  - Classification
  - Regression
- Unsupervised learning
  - Clustering
- Bias-Variance Tradeoff
- Occam’s Razor

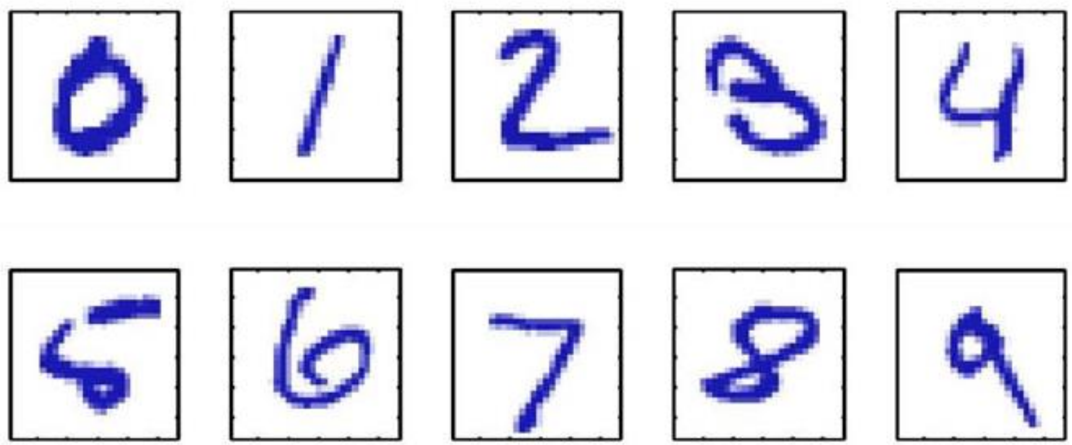
Slides adapted from

- A. Zisserman, University of Oxford, UK
- S. Ullman, T. Poggio, D. Harari, D. Zysman, D Seibert, MIT
- D. Sontag, MIT
- Figures from “An Introduction to Statistical Learning”, James et al.

Introduction

- What is Machine Learning?
  - Subset of AI
  - Design algorithms that learn from real data and can automate critical tasks
- When can it be applied?
  - It cannot solve any problem!
  - When task can be expressed as learning task
  - When high-quality data is available
    - Labeled data (by human experts) is preferable!
  - When some error is acceptable (can rarely achieve 100% accuracy)
    - Example: recommendation system, advertisement engine

Example 1
Handwritten digit recognition



Images are 28 x 28 pixels

Represent input image as a vector $\mathbf{x} \in \mathbb{R}^{784}$

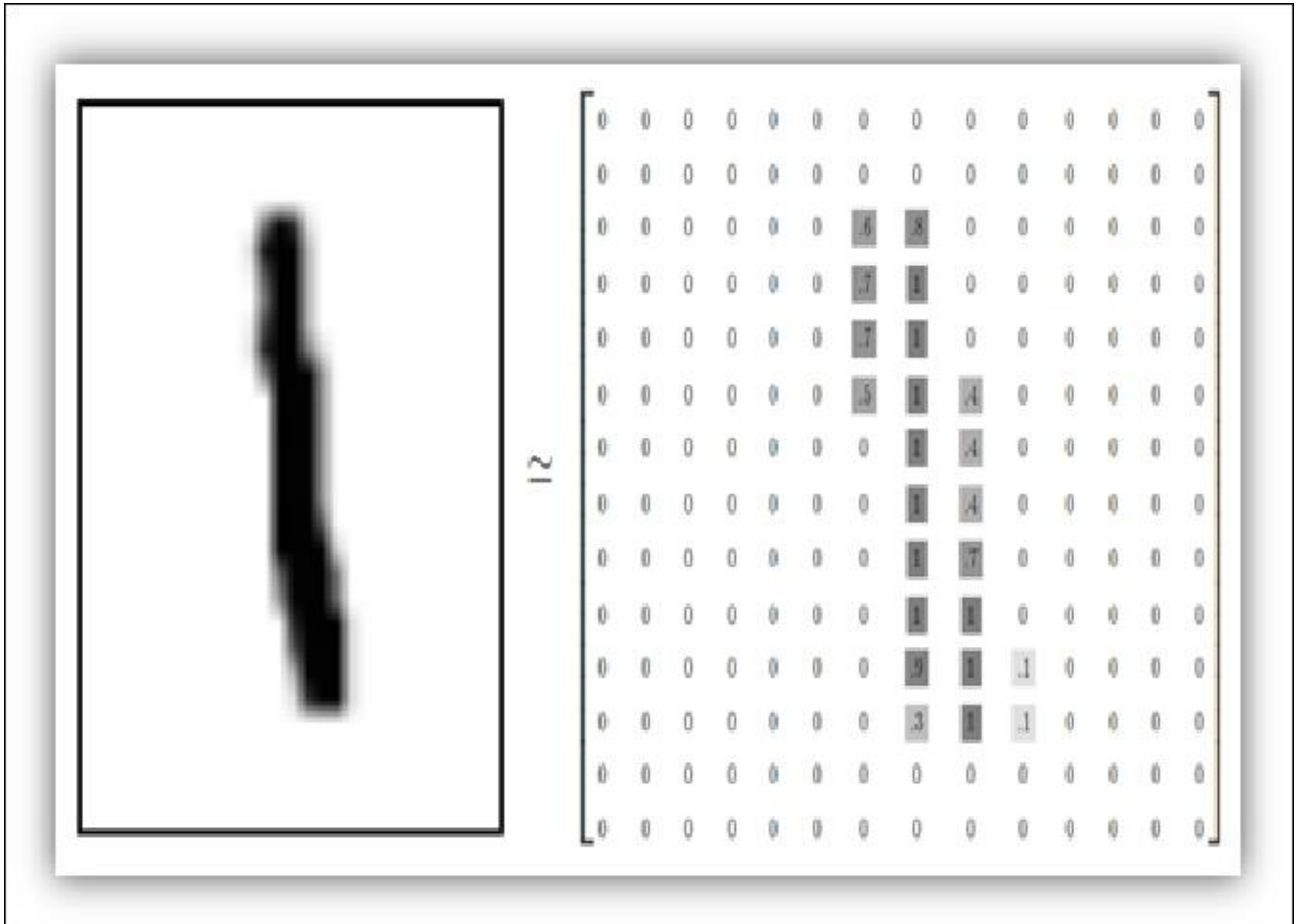
Learn a classifier $f(\mathbf{x})$ such that,

$$f : \mathbf{x} \to \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$$

MNIST dataset: Predict the digit

Multi-class classifier

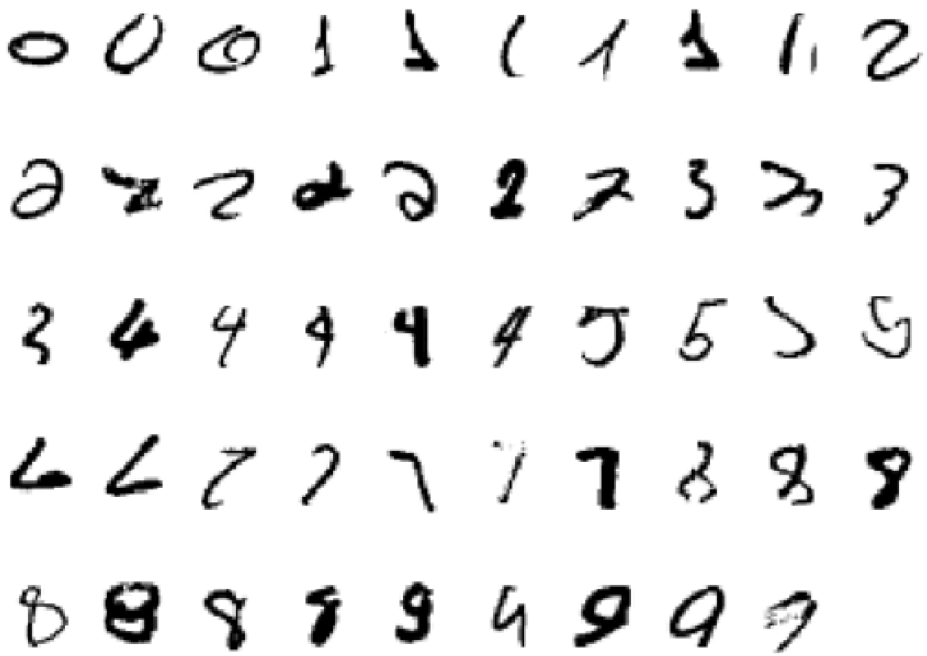
Data Representation



Model the problem

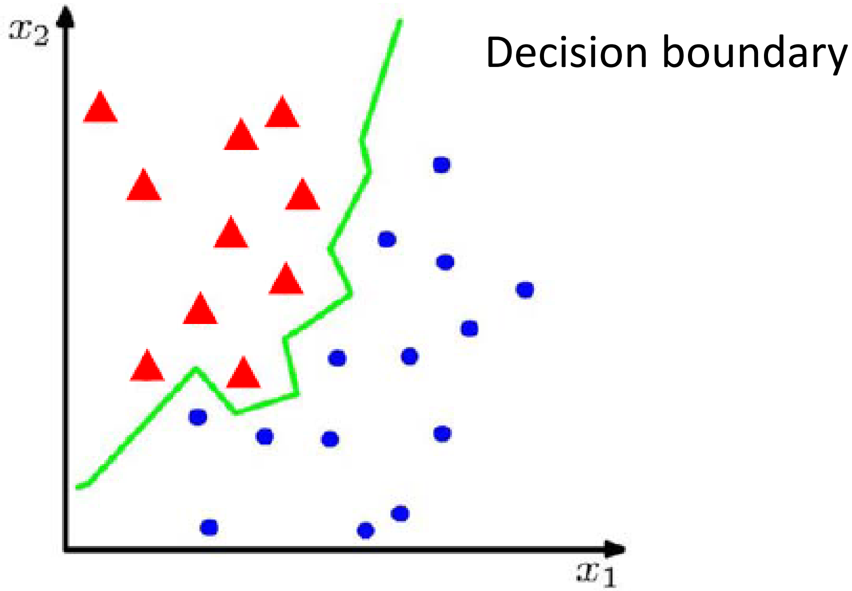
As a supervised classification problem

Start with training data, e.g. 6000 examples of each digit



- Can achieve testing error of 0.4%
- One of first commercial and widely used ML systems (for zip codes & checks)

Classification



- Suppose we are given a training set of N observations

$x^{(1)}, \dots, x^{(N)}$ and $y^{(1)}, \dots, y^{(N)} \in \{0,1\}$ Binary

- Classification problem is to estimate f(x) from this data such that

$$f(x^{(i)}) = y^{(i)}$$

Extended to multi-class classification

- Handwritten digit recognition

Supervised Learning: Classification

Training

Data → Pre-processing → Feature extraction → Learning model

Labeled

$x^{(i)}, y^{(i)} \in \{0,1\}$

Normalization

Feature Selection

Classification

$f(x)$

Testing

New data → Learning model → Predictions $y' = f(x') \in \{0,1\}$

Unlabeled

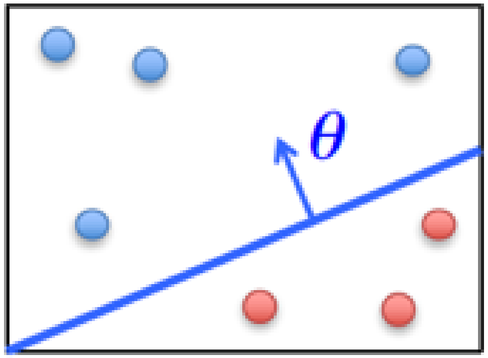
x'

$f(x)$

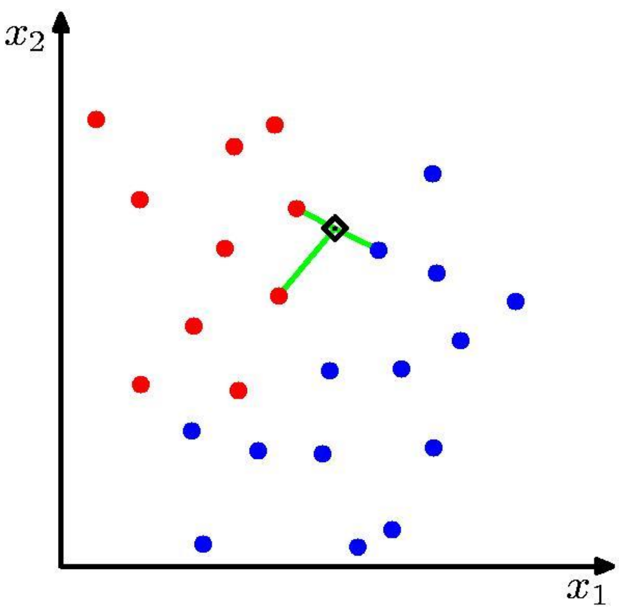
Positive
Negative

Classification

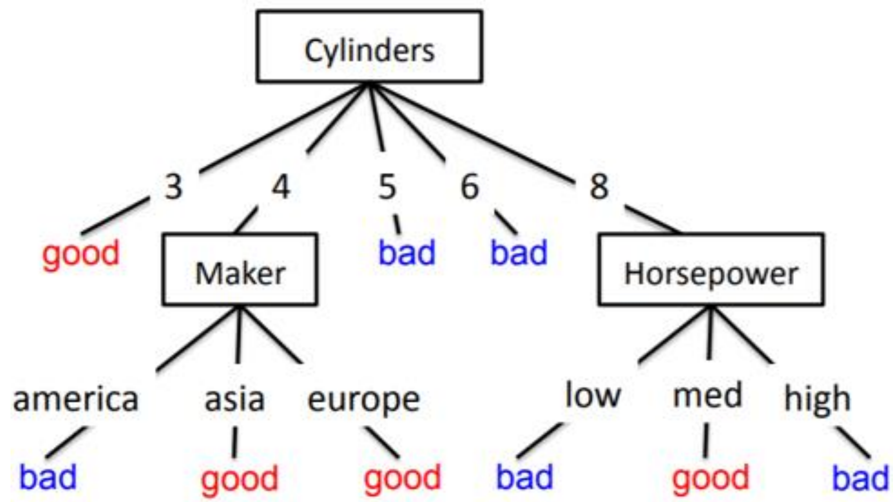
Example Classifiers



Linear classifiers: logistic regression, SVM, LDA



Nearest Neighbors (kNN)



Decision trees