

# DS 4400

## Machine Learning and Data Mining I Spring 2021

Alina Oprea

Associate Professor

Khoury College of Computer Science

Northeastern University

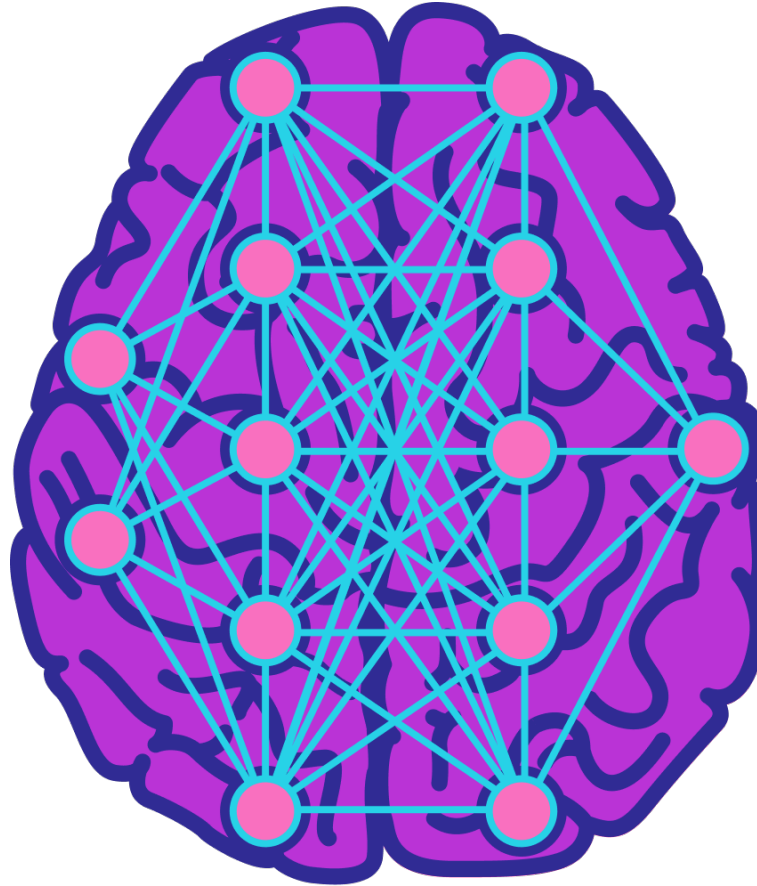
January 19 2021

# Recording

The class will be recorded and the recordings  
made available online

To opt out: send a message in the Chat

# Welcome to DS 4400!



## Machine Learning and Data Mining I

# Introduction

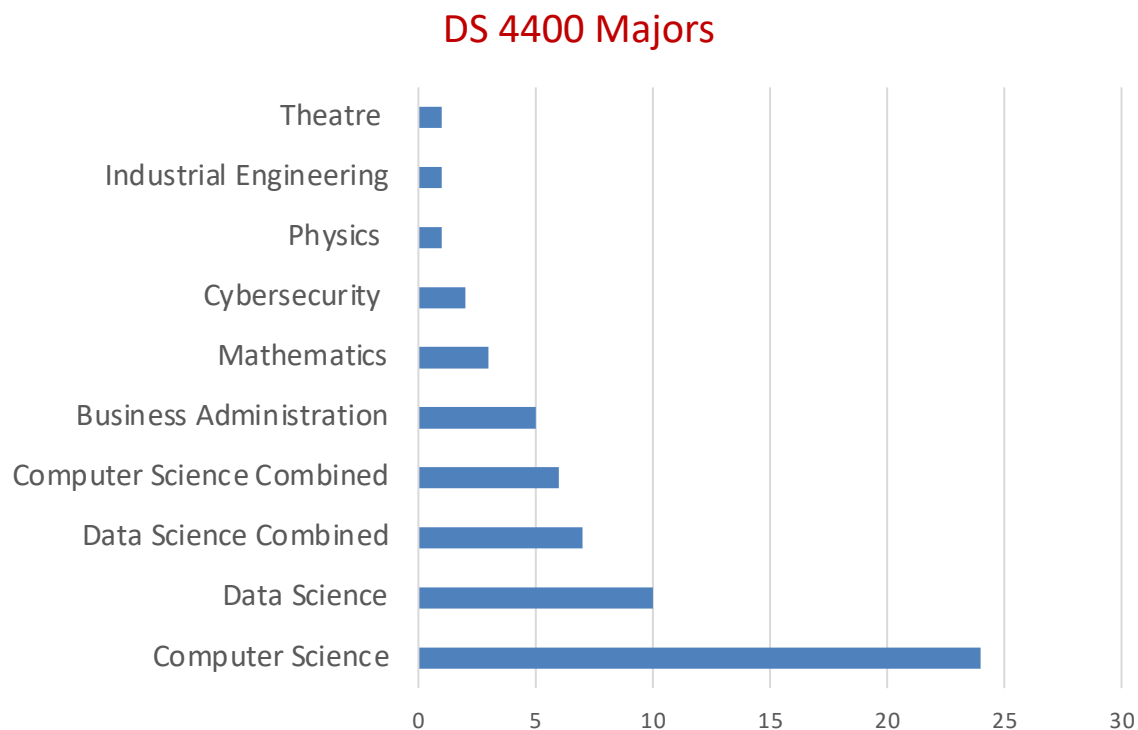
- **Ph.D. at CMU**
  - Research in applied cryptography, data security, and cryptographic file systems
- **RSA Laboratories**
  - Cloud security, applied cryptography, game theory for security
  - ML/AI in security
- **NEU Khoury College – since Fall 2016**
  - NDS2 Lab part of the Cybersecurity and Privacy Institute
  - Machine learning for security applications: attack detection, IoT, connected cars, collaborative defenses
  - Adversarial machine learning: study the vulnerabilities of ML in face of attacks and design defenses
  - Privacy in machine learning

# TA Introduction




- Omkar Reddy Gojala
  - MS in Data Science
  - Research experience in the Barabasi Lab
- Prabal Malviya
  - MS in Data Science
  - TA for Foundations of Data Science
  - Data science co-op at Danfoss
- Saurabh Nitin Parkar
  - MS in Data Science
  - Co-op at Broad Institute

# DS 4400 Class

- Enrollment of 60
- Diverse majors

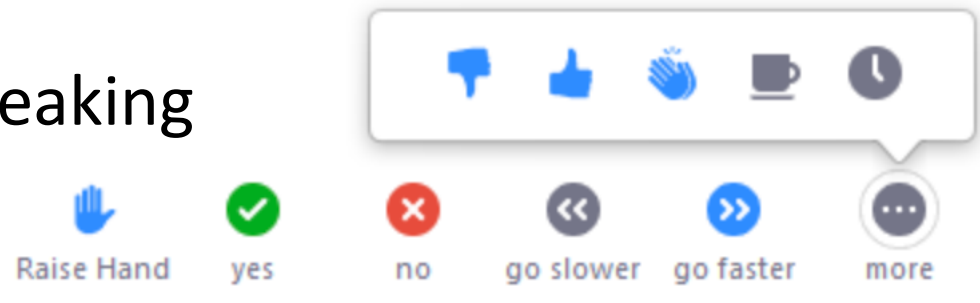


# Course Information

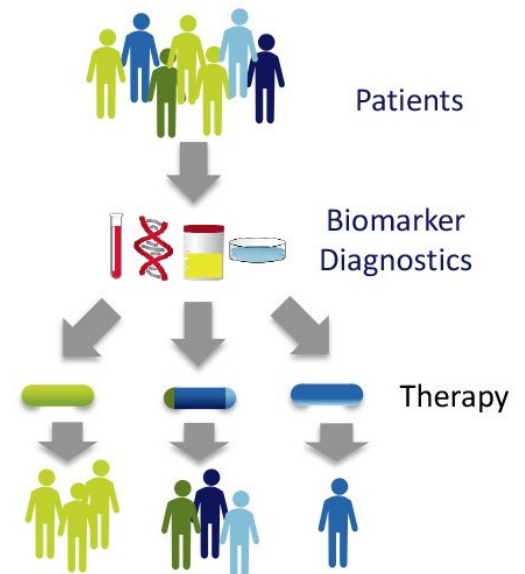
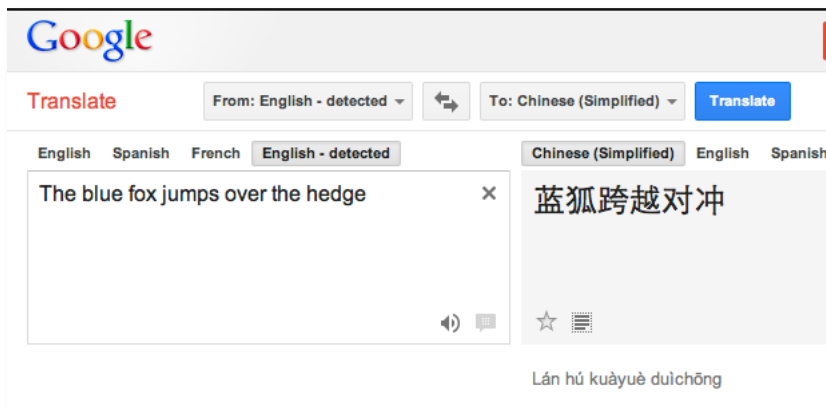
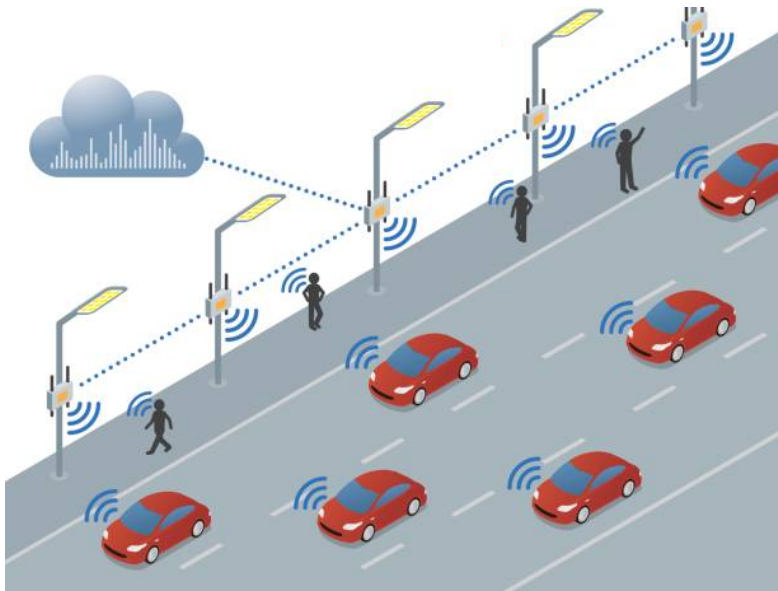
- Website:  
[www.ccs.neu.edu/home/alina/classes/Spring2021](http://www.ccs.neu.edu/home/alina/classes/Spring2021)
- Canvas: <https://canvas.northeastern.edu>   
canvas
- Gradescope: [gradescope.com](https://gradescope.com)  gradescope
- Communication: [piazza.com](https://piazza.com) 
- E-mail:
  - [a.oprea@northeastern.edu](mailto:a.oprea@northeastern.edu)
  - [gojala.o@northeastern.edu](mailto:gojala.o@northeastern.edu)
  - [malviya.p@northeastern.edu](mailto:malviya.p@northeastern.edu)
  - [parkar.s@northeastern.edu](mailto:parkar.s@northeastern.edu)

# Online Classes

- Zoom conference call for class lectures
- Log in at [northeastern.zoom.us](https://northeastern.zoom.us)
  - Upload a profile picture
  - Turn video on
  - Mute when not speaking
- Provide feedback
- To ask questions:
  - Raise hand
  - Use chat
- Recording will be posted online



# Machine Learning is Everywhere



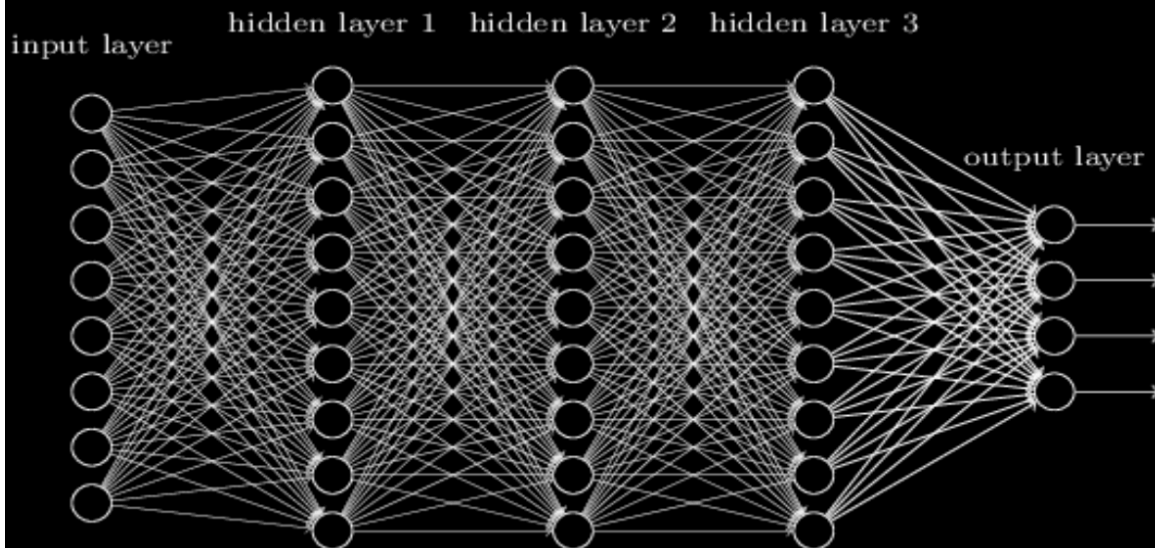
# Short History

- Legendre and Gauss – linear regression, 1805
  - Astronomy applications
- Probabilistic models
  - Bayes and Laplace - Bayes Theorem, 1812
  - Markov chains, 1913
- Fisher – linear discriminant analysis for classification, 1936
  - Logistic regression, 1940
- Widrow and Hoff ADALINE neural network, 1959
- Nelder, Wedderburn, generalized linear models, 1970
- “AI winter”, limitations of perceptron and linear models, 1970
- Breiman, Friedman, Olshen, Stone, decision trees (non-linear models), 1980
- Cortes and Vapnik, SVM with kernels, 1990
- IBM Deep Blue beats Kasparov at chess, 1996
- Geoffrey Hinton, Deep learning, back propagation, 2006
- C. Szegedy: Adversarial manipulation of image classification, 2013

# Deep Learning

Neural networks return and excel at image recognition, speech recognition, ...

The 2018 Turing award was given to Yoshua Bengio, Geoff Hinton, and Yann LeCun.



# Safety Concerns of AI

- Ethics and fairness of AI
  - Everyone is treated fairly
  - Robots will not perform harmful actions
  - Can the technology be used for nefarious purposes?
- Economic concerns
  - Might automate / displace some type of jobs in manufacturing, transportation, etc.
- Adversarial ML
  - ML can be manipulated
  - Small change in input results in different prediction

# Secure and Robust ML

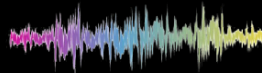
## Image Recognition

Misreading traffic signs  
(Eykholt et al)



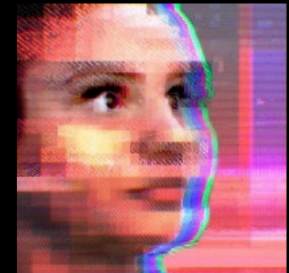
## Speech recognition

Hide commands in  
noise (Carlini & Wagner)



## Poisoning Attacks

Tay (chat bot) became  
inflammatory in 16 hr.



How to create safe and robust machine  
learning?

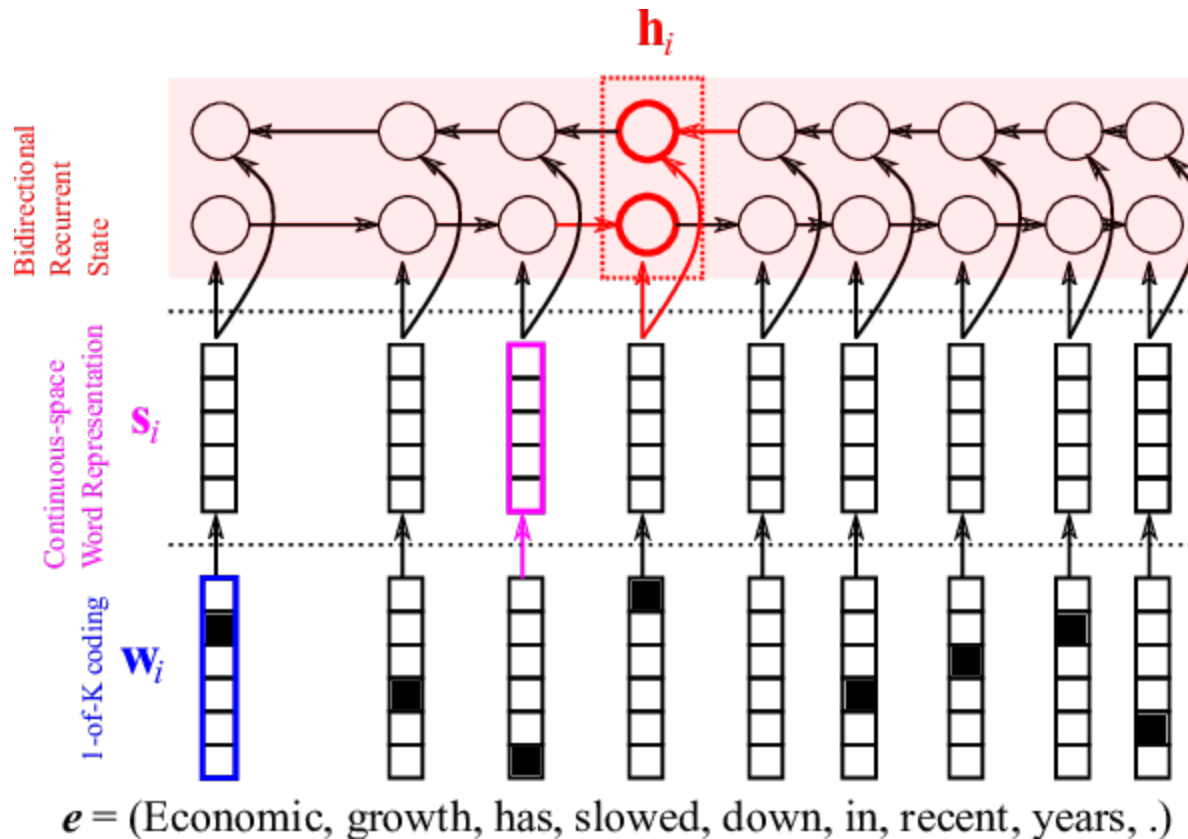
# Discussion

- Discuss most exciting ML applications
- What are some of the concerns when using ML in the real world?

# Applications of ML

- Healthcare
- Vision
- NLP
- Speech recognition
- Self-driving cars
- Stock market analysis
- Recommendations
- Sentiment analysis
- Human behavior
- Quality of life
- Business
- Sports
- Bots / chatbots
- Science / engineering
- Bioinformatics
- Precision medicine
- Unsupervised learning
- Reinforcement learning

# Natural Language Processing (NLP)

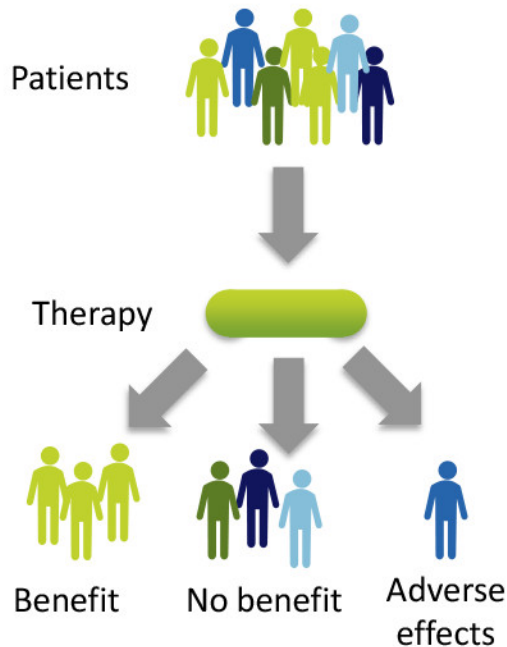


- Understand language semantics
- Real-time translation, speech recognition, question answering
- Large generative language models: BERT, GPT-2, GPT-3

# Personalized medicine

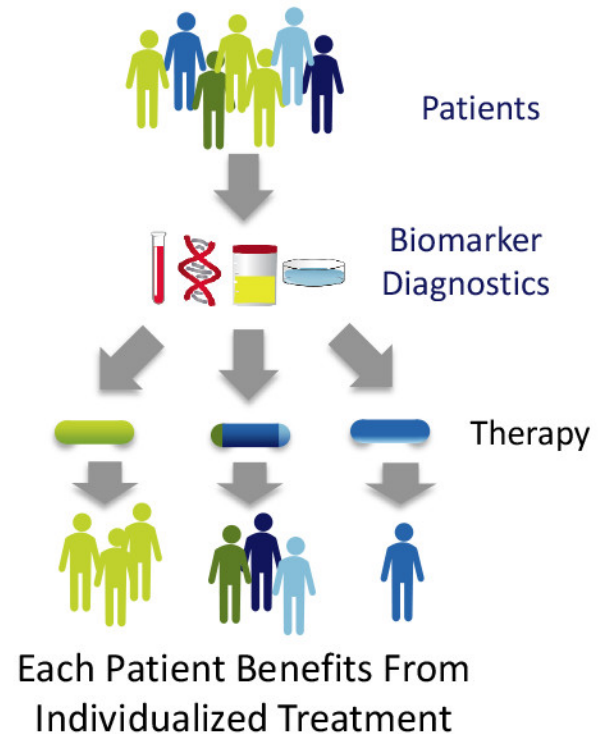
## Without Personalized Medicine:

Some Benefit, Some Do Not



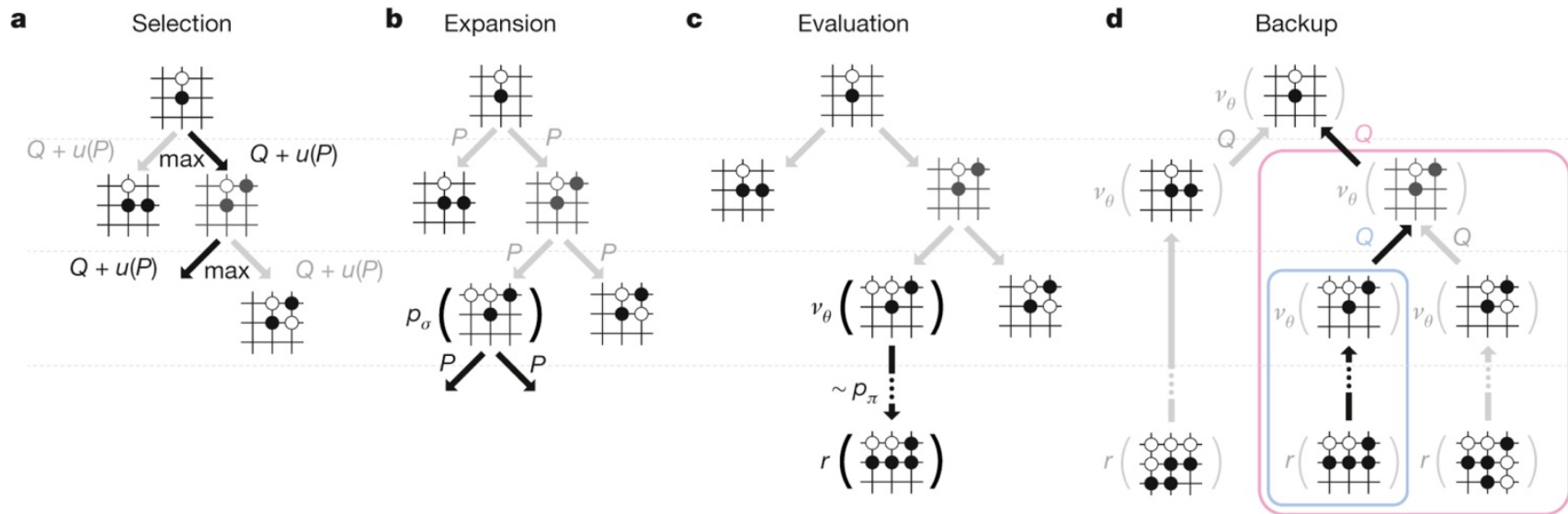
## With Personalized Medicine:

Each Patient Receives the Right Medicine For Them



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

# Playing games



- AlphaGo: DeepMind beats world champion in 2015
- Interestingly, it discovered new, unknown strategies
- Go is the most challenging game for AI
- Algorithms based on deep reinforcement learning

# 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 main machine learning tasks
  - Supervised learning vs unsupervised learning
  - Classification vs Regression
- Study most well-known algorithms
  - Regression (linear regression, spline regression)
  - Classification (SVM, decision trees, Naïve Bayes, ensembles, etc.)
  - Deep learning (different neural network architectures)
- Learn the theory and foundation behind ML algorithms and learn to apply them to real datasets
- Learn about security challenges of ML and ethical issues
  - Introduction to adversarial ML

<http://www.ccs.neu.edu/home/alina/classes/Spring2021/>

# Class Outline

- Introduction – 1 week
  - Probability and linear algebra review
- Linear regression and regularization – 2 weeks
- Classification - 5 weeks
  - Linear classifiers: logistic regression, LDA,
  - Non-linear: kNN, decision trees, SVM, Naïve Bayes
  - Ensembles: random forest, boosting
  - Model selection, regularization, cross validation
- Neural networks and deep learning – 2 weeks
  - Back-propagation, gradient descent
  - NN architectures (feed-forward, convolutional, recurrent)
- Ethics of AI – 1 lecture
- 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](#)

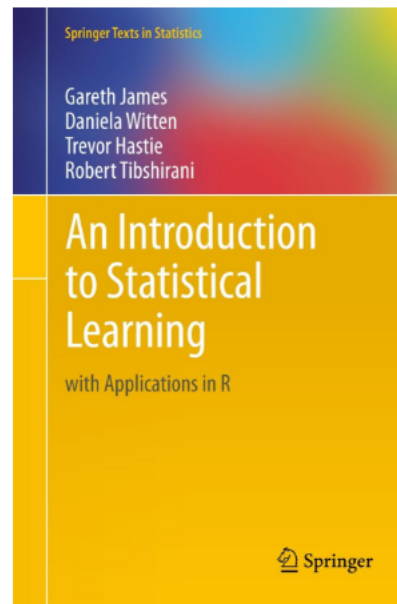
[Data Sets and Figures](#)

[ISLR Package](#)

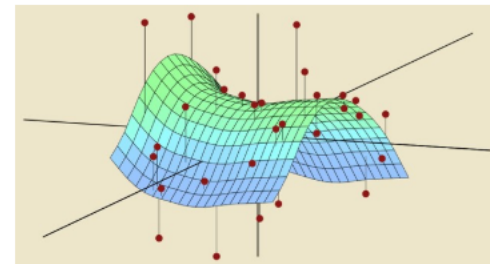
[Get the Book](#)

[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.
- A. Zhang, Z. Lipton, and A. Smola. [Dive into Deep Learning](#)
- Lecture notes by Andrew Ng from Stanford

# Policies

- **Instructors**
  - Alina Oprea
  - TAs: Omkar Reddy Gojala, Prabal Malviya, Saurabh Nitin Parkar
- **Schedule**
  - Tue 11:45am – 1:25pm, Thu 2:50-4:30pm EST
  - Shillman Hall 320 and Zoom lectures
  - Office hours (Zoom):
    - Alina: Tuesday 4:30-5:30pm; Thursday 4:30 – 5:30 pm
    - Omkar: Monday and Wednesday 3:00-4:00pm;
    - Prabal: Monday and Thursday 12:00-1:00pm
    - Saurabh: Friday 10am-12pm
    - Links on Canvas under “Syllabus”
- **Online resources**
  - Slides / recordings will be posted after each lecture for 48 hours
  - Use Piazza for questions
  - Canvas as course management system

# Grading

- **Assignments – 25%**
  - 4-5 assignments and programming exercises based on studied material in class
- **Final project – 30%**
  - Select your own project based on public dataset
  - Submit short project proposal and milestone
  - Presentation at end of class (10 min) and written report
  - Team of 2 students
- **Midterm Exam –20%**
  - Tentative date: Tuesday, March 2
- **Final Exam – 20%**
  - Tentative date: Tuesday, April 6
- **Class participation – 5%**
  - Participate in class discussion/Zoom and on Piazza
  - Pop up quizzes

# Academic Integrity

- Homework is done individually!
- Final project is done in the team!
- Rules
  - Can discuss with colleagues or instructors
  - 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 CHEATING 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/>