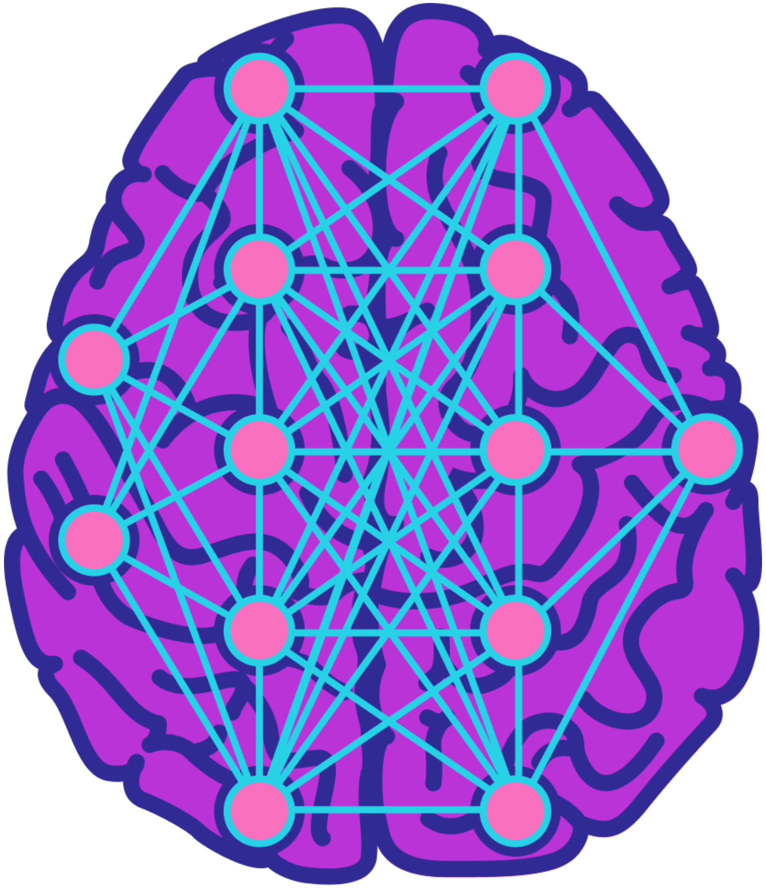
# DS 4400

## Machine Learning and Data Mining I
## Spring 2022

Alina Oprea
Associate Professor
Khoury College of Computer Science
Northeastern University

January 19 2022

# Welcome to DS 4400!



Machine Learning and Data Mining I

# DS 4400 Class

- Enrollment of 110
- Diverse majors
  - Computer Science
  - Data Science
  - Cybersecurity
  - Economics
  - Neuroscience
  - Bioengineering
  - Finance
  - Math
  - Joint majors: DS/Psychology, DS/Business, DS/Math, DS/Biology, etc.

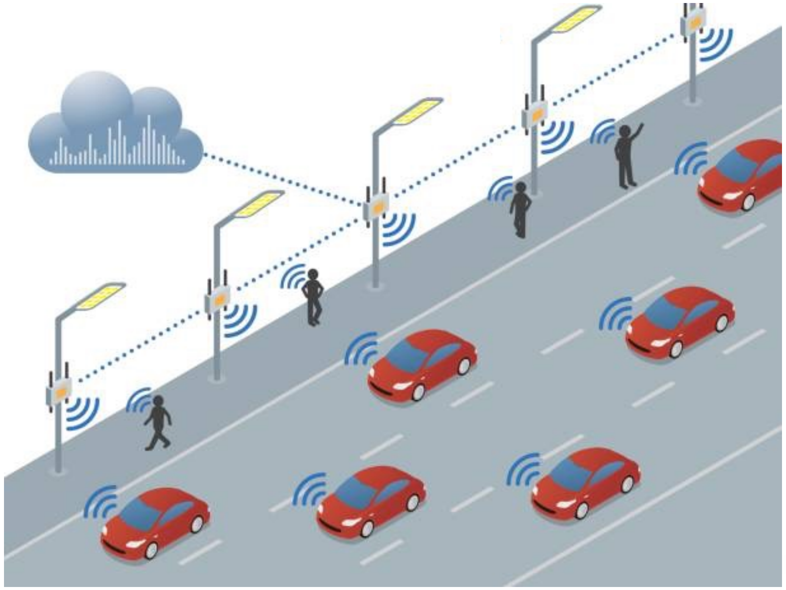
# Introduction

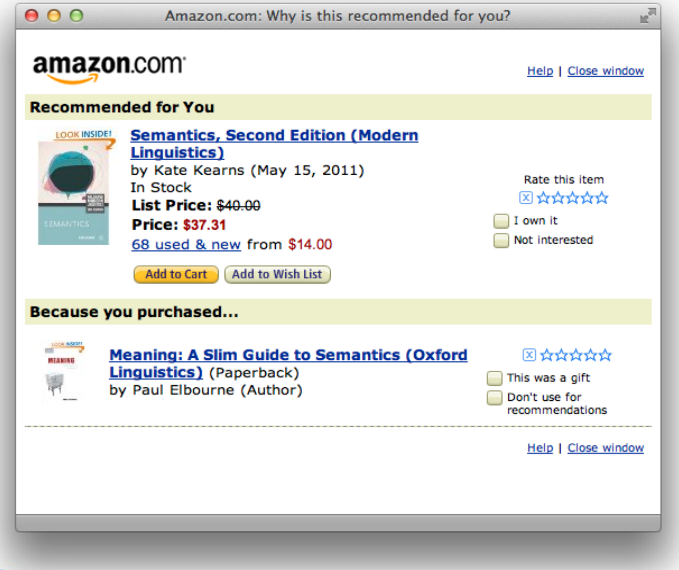
- Ph.D. at CMU
  - Research in applied cryptography, data security, and cryptographic file systems
- RSA Laboratories
  - Cloud security, applied cryptography, game theory for security
  - ML/AI in security
- NEU Khoury College – since Fall 2016
  - NDS2 Lab part of the Cybersecurity and Privacy Institute
  - Affiliated with the Experential AI 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

- Sri Harika Cherukuri cherukuri.s@northeastern.edu
  - 1st year MS student in data science
- Nathaniel Hofmann hofmann.n@northeastern.edu
  - 5th year undergrad in CS, minor in Math
- Jake Horban horban.y@northeastern.edu
  - 3rd year undergrad in DS, minor in Math/Economics
- Noah Lee lee.no@northeastern.edu
  - 3rd year undergrad in CS
- Xuyang Li li.xuya@northeastern.edu
  - Undergrad in Math / Combined DS and biochemistry, minor in physics
- Talha Ongun ongun.t@northeastern.edu
  - 5th year PhD student at Khoury

# Machine Learning is Everywhere





Amazon.com: Why is this recommended for you?

amazon.com

Help | Close window

**Recommended for You**

**Semantics, Second Edition (Modern Linguistics)**
by Kate Kearns (May 15, 2011)
In Stock
**List Price:** ~~$40.00~~
**Price:** $37.31
68 used & new from $14.00

Add to Cart | Add to Wish List

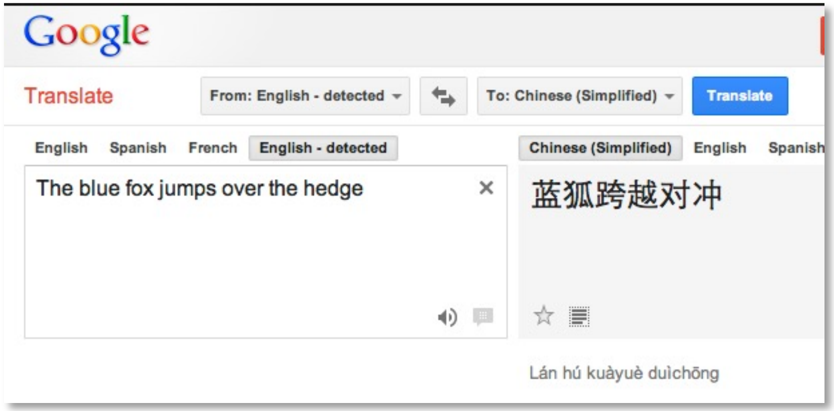
Rate this item
☒ ☆☆☆☆☆
□ I own it
□ Not interested

**Because you purchased...**

**Meaning: A Slim Guide to Semantics (Oxford Linguistics)** (Paperback)
by Paul Elbourne (Author)

☒ ☆☆☆☆☆
□ This was a gift
□ Don't use for recommendations

Help | Close window



Google

Translate

From: English - detected | To: Chinese (Simplified) | Translate

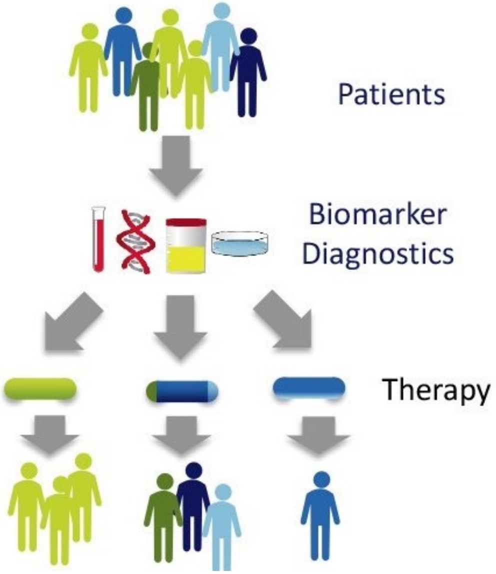
English | Spanish | French | English - detected

Chinese (Simplified) | English | Spanish

The blue fox jumps over the hedge

蓝狐跨越对冲

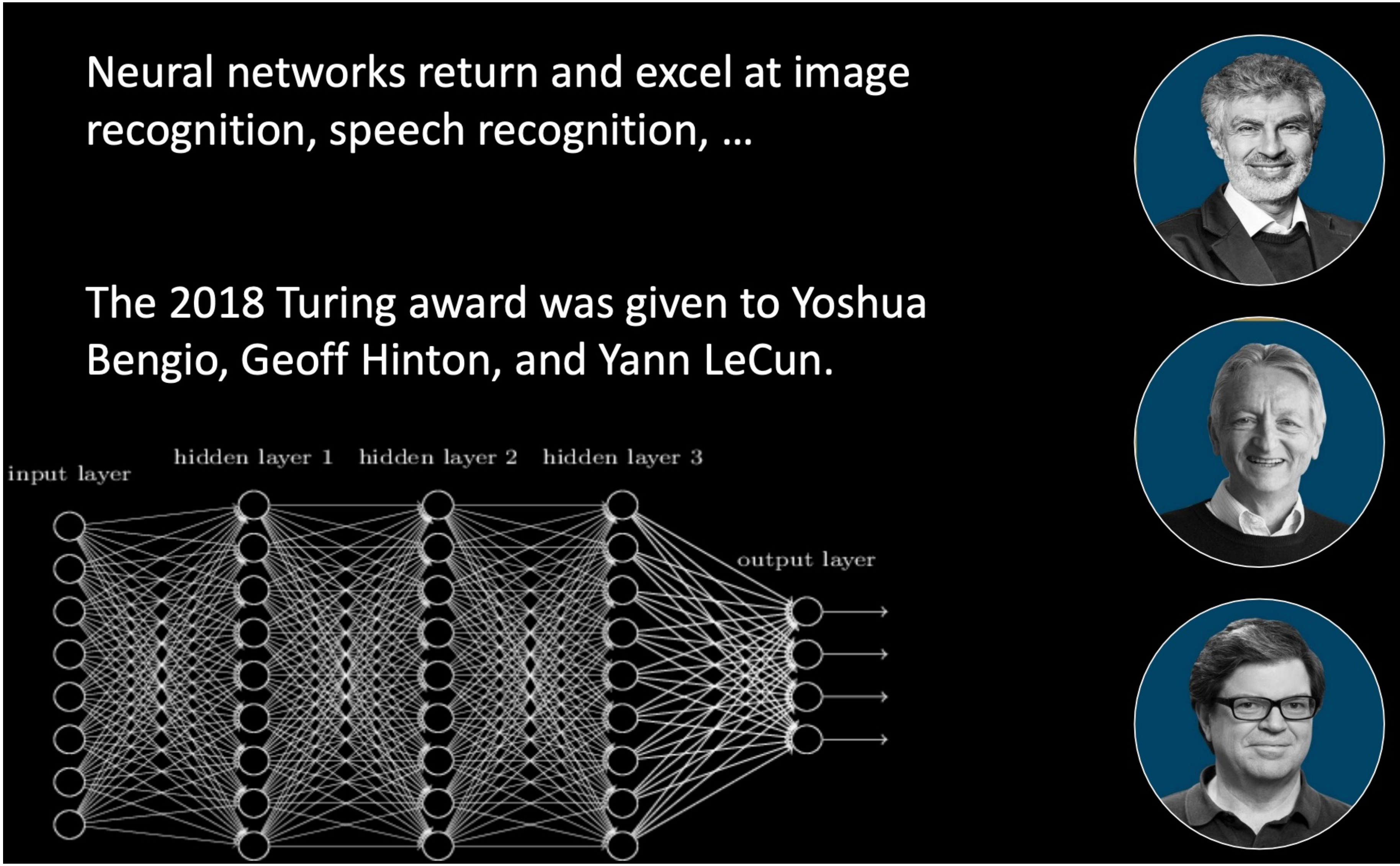
Lán hú kuàyuè duìchōng



# Short History

- Legendre and Gauss – linear regression / least squares, 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. Szedegy: 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.

STOP
STOP
STOP
12:45

How to create safe and robust machine learning?

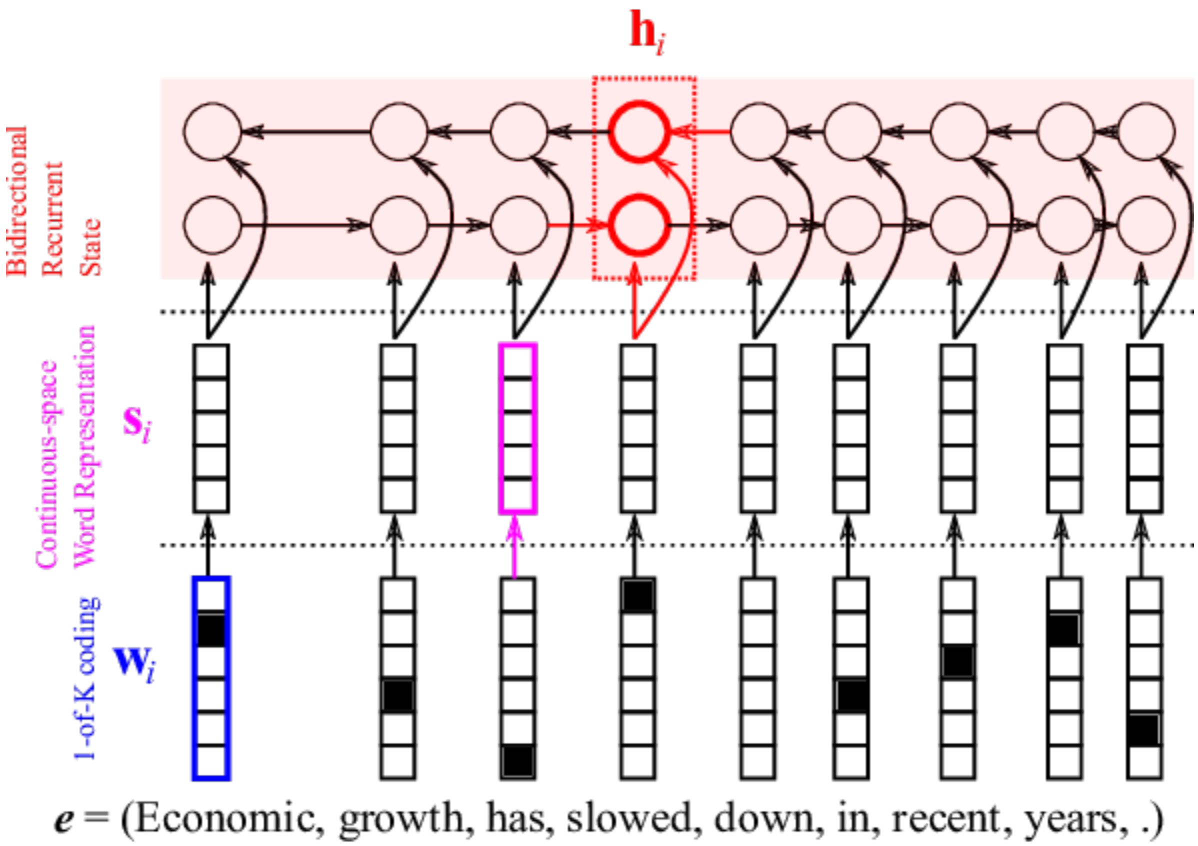
# Discussion

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

# Applications of ML

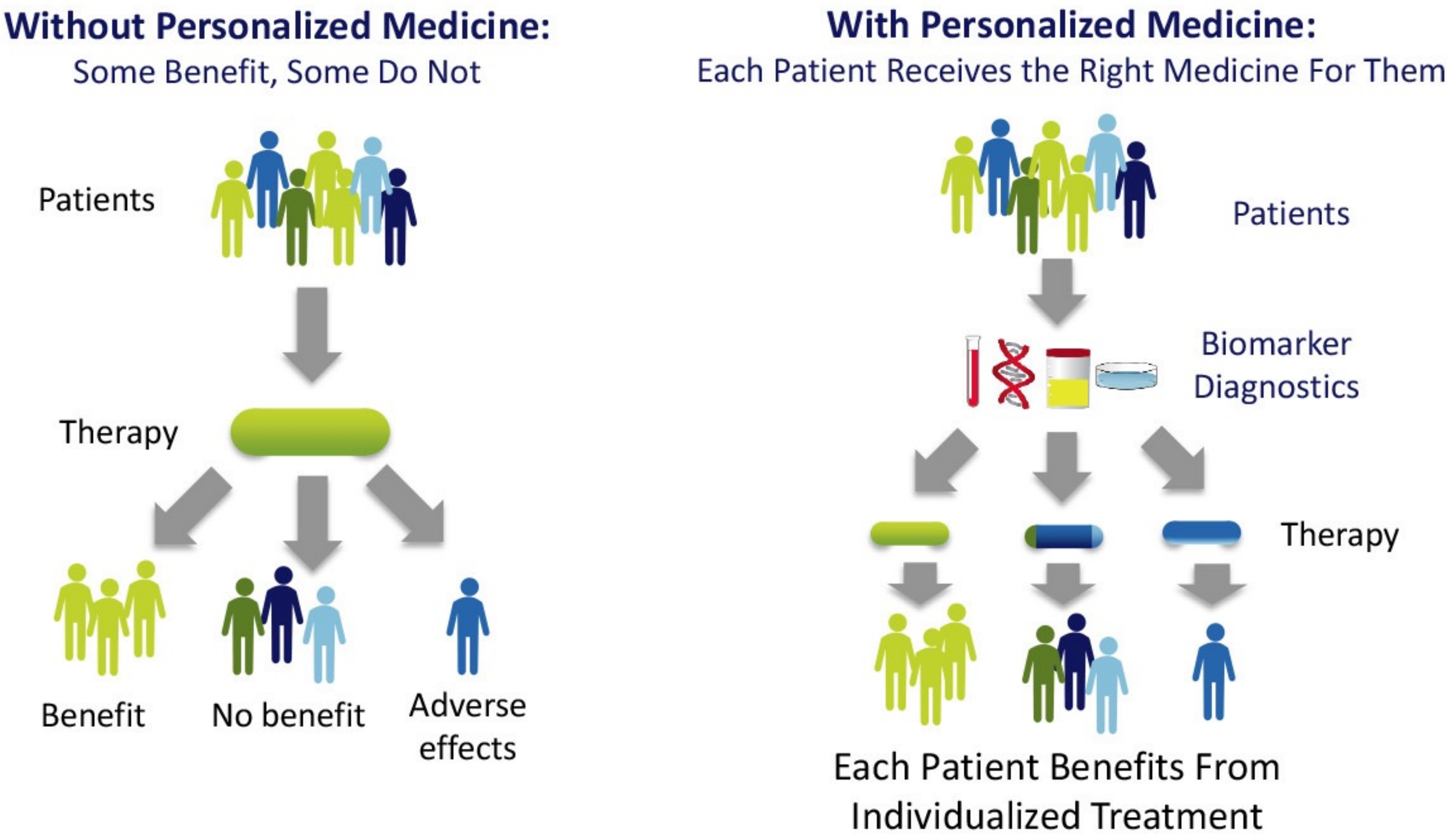
- Healthcare
- Vision
- NLP
- Speech recognition
- Self-driving cars
- Stock market analysis
- Recommendations
- Sentiment analysis
- Human behavior
- Quality of life
- Business
- Sports
- Bots / chatbots
- Science / engineering
- Bioinformatics
- Precision medicine
- Unsupervised learning
- Reinforcement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

# Natural Language Processing (NLP)


$\mathbf{h}_i$
Bidirectional Recurrent State
Continuous-space Word Representation
$\mathbf{s}_i$
1-of-K coding
$\mathbf{w}_i$
$e$ = (Economic, growth, has, slowed, down, in, recent, years, .)

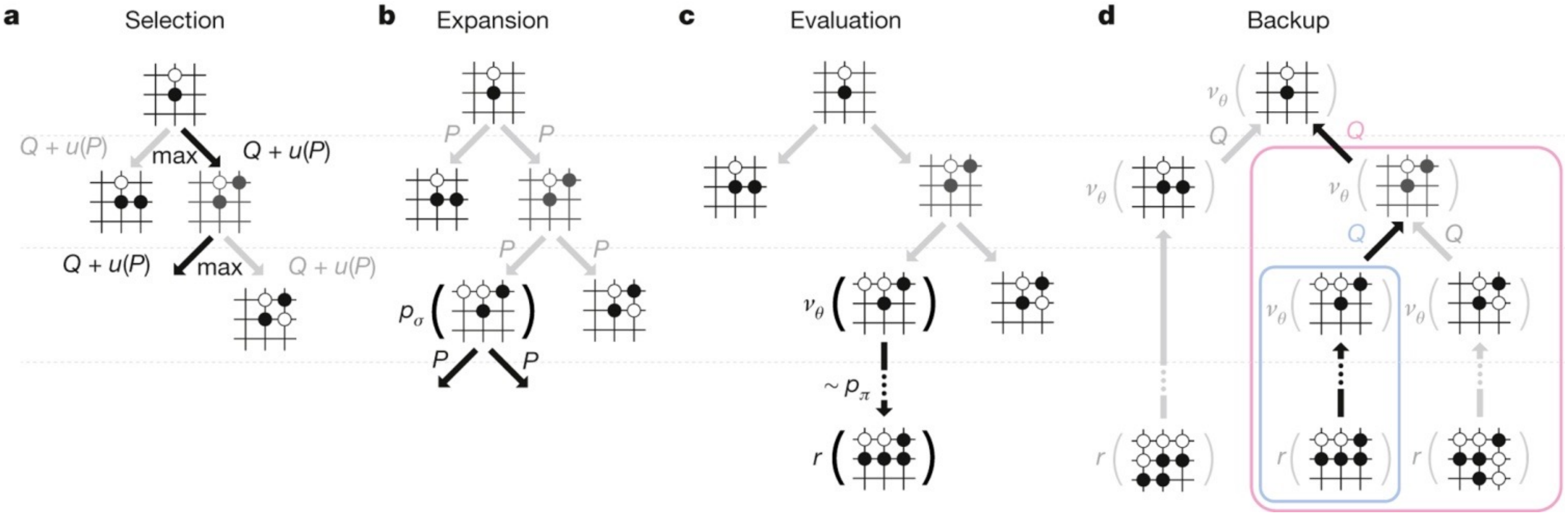
- Understand language semantics
- Real-time translation, speech recognition, question answering
- Large generative language models: BERT, GPT-2, GPT-3

# Personalized medicine

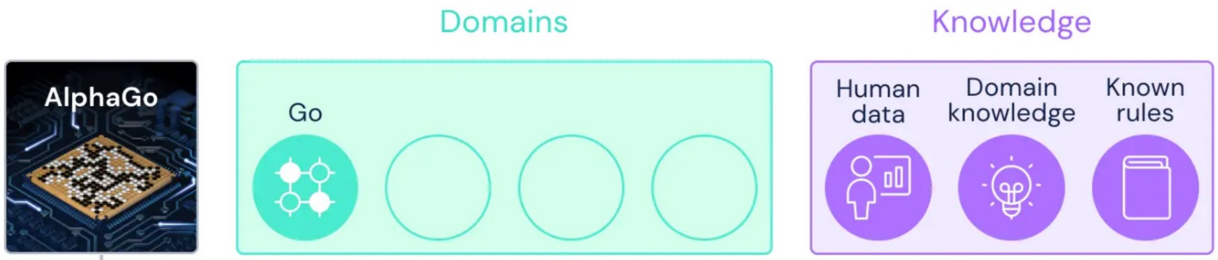

Without Personalized Medicine:
Some Benefit, Some Do Not
Patients
Therapy
Benefit
No benefit
Adverse effects
With Personalized Medicine:
Each Patient Receives the Right Medicine For Them
Patients
Biomarker Diagnostics
Therapy
Each Patient Benefits From Individualized Treatment

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

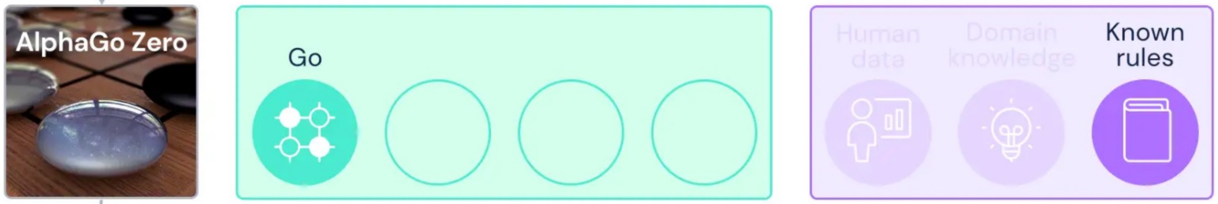
# Playing games



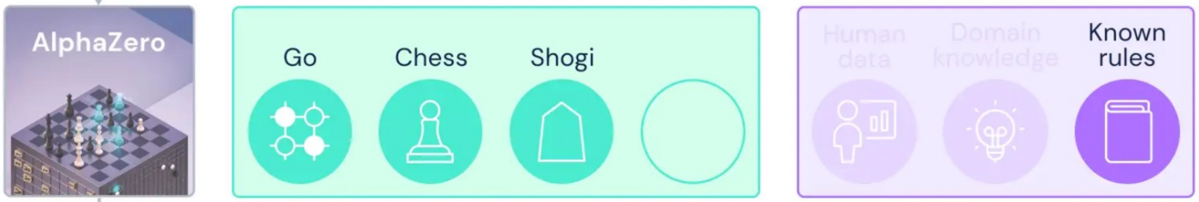
- AlphaGo: DeepMind beats world champion in 2016
- Interestingly, it discovered new, unknown strategies
- Go is the most challenging game for AI
- Algorithms based on deep reinforcement learning



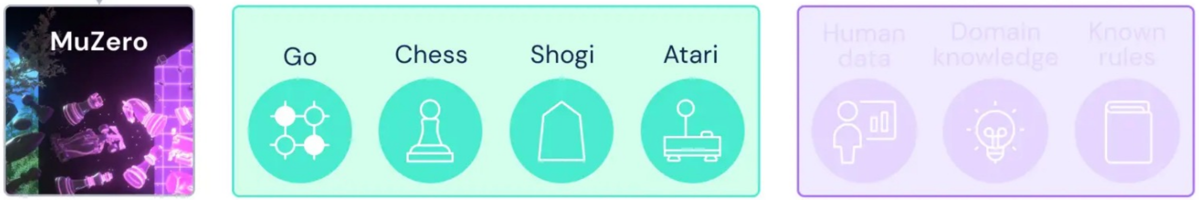
**AlphaGo** becomes the first program to master Go using neural networks and tree search
(Jan 2016, Nature)



**AlphaGo Zero** learns to play completely on its own, without human knowledge
(Oct 2017, Nature)



**AlphaZero** masters three perfect information games using a single algorithm for all games
(Dec 2018, Science)



**MuZero** learns the rules of the game, allowing it to also master environments with unknown dynamics.
(Dec 2020, Nature)

# DS-4400

- What is *machine learning*?
  - The development of computer systems that are able to learn and adapt without following explicit instructions, by using algorithms and statistical models to analyze and draw inferences from patterns in data
  - Design predictive algorithms that learn from data
  - Subset of Artificial Intelligence (AI): The study of "intelligent agents" that perceive their environment and take actions that maximize their chance of achieving their goals
- Machine learning is currently very successful in:
  - Machine translation
  - Voice assistants
  - Recommendation systems
  - Image recognition
- 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
  - Focus on supervised learning
- 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/Spring2022/

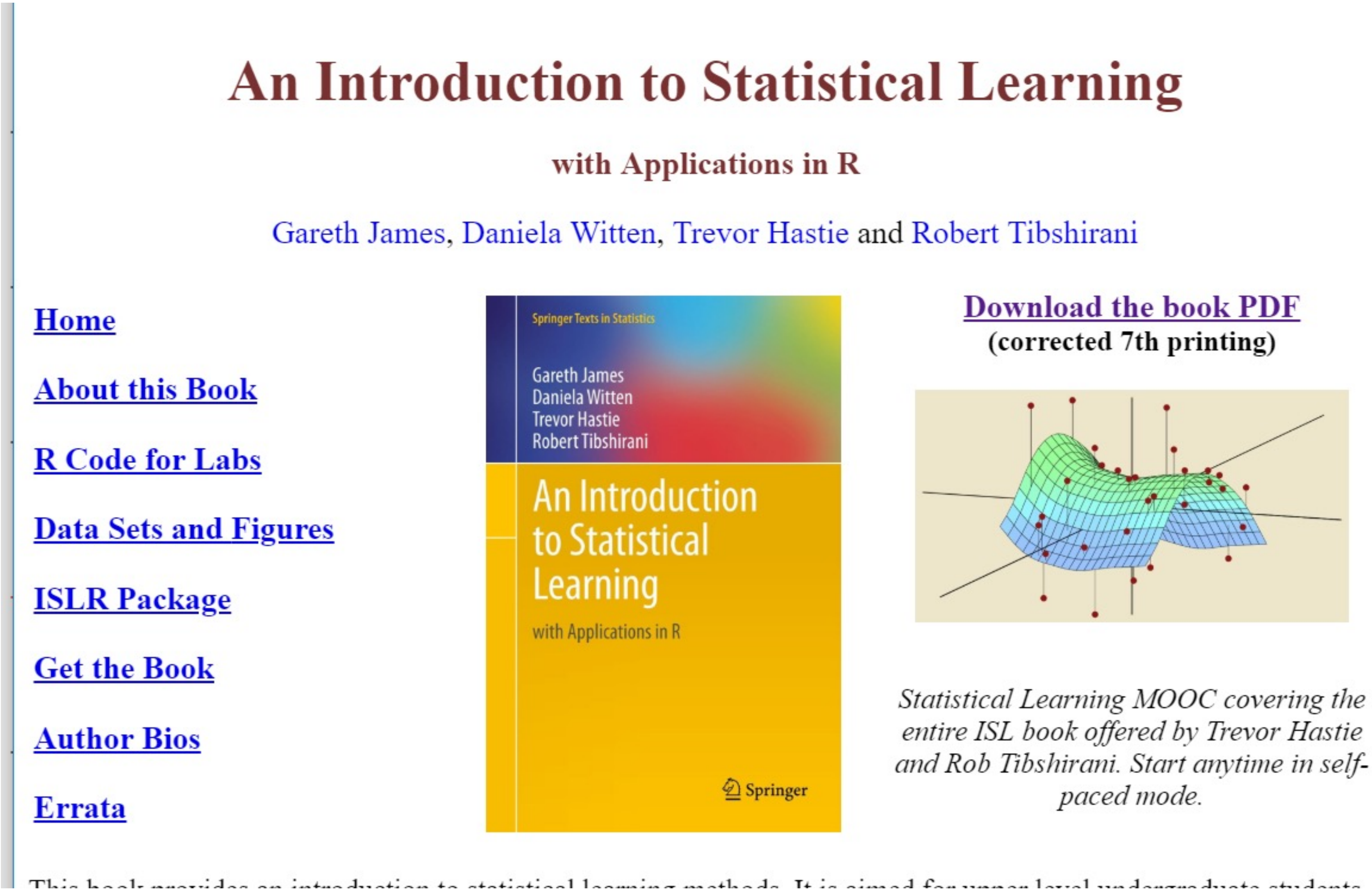
# Class Outline

- 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 – 2 lectures
- Adversarial ML – 1 lecture
  - Security of ML at testing and training time

# Course Information

- Website: http://www.ccs.neu.edu/home/alina/classes/Spring2022
- Canvas: https://canvas.northeastern.edu
- Gradescope: gradescope.com
- Communication: piazza.com

# 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

# Schedule

- Schedule
  - Mon, Wed 2:50-4:30pm ET, Richards Hall 236
  - Office hours (Zoom), links on Canvas:
    - Alina: Mon and Wed after class 4:45-5:45pm (Mon session reserved for students who could not attend class)
    - Sri Harika Cherukuri: Mon 12-1:30pm
    - Nathaniel Hofmann: Tue 4:30-6pm
    - Jake Horban: Wed 12-1:30pm
    - Noah Lee: Tue 9-10:30am
    - Xuyang Li: Thu 4:30-6pm
    - Talha Ongun: Fri 4-5:30pm
- Online resources
  - Slides and lecture notes will be posted after each lecture
  - Use Piazza for questions

# Policies

- 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
  - We will use Gradescope for submitting assignments
  - No need to email for late days

# 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
  - Record presentation (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%
  - Pop up quizzes

# Assignments

- Several theoretical questions and many programming exercises
- Language
  - Python
  - Jupyter notebooks recommended
  - Will share some numpy and panda tutorials
- Submission
  - Submit PDF report
  - Includes all the results, as well as link to code

# Final project

- Goal: work on a larger data science project
  - Build your portfolio and increase your experience
- Requirements
  - Large dataset: at least 20,000 records (public source)
  - Not recommended to collect your own data
  - Pick application of interest
  - We will also provide a list of projects and datasets
  - Experiment with at least 4 ML models
  - Perform in-depth analysis (which features contribute mostly to prediction, which model performs best, explain results)
  - Teams of 2 students, will have a TA assigned
- Computational resources: NEU Discovery cluster, Google cloud, AWS, Google collab
- Timeline
  - Proposal: mid class; milestone 3 weeks after (Instructors will provide early feedback)
  - Final presentation (recorded video) and report (6-8 pages)

# 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 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/