STA 314: Statistical Methods for Machine Learning I

Logistics and why this course

Xin Bing

Department of Statistical Sciences
University of Toronto

About this course

- This course is a broad introduction to machine learning from a statistical perspective (aka statistical learning). We put emphasis on intuition and basic mathematical derivations of how (and why) popular machine learning methods work.

- We will focus on understanding methodology rather than implementing complicated machine learning algorithms or delving into deep theory.

- You will learn examples of applying popular machine learning methods to real data sets in R.

About this course

We cover two types of learning problems:

- Supervised learning (80%)
  - Regression
  - Classification
- Unsupervised learning (20%)
  - Dimension reduction
  - Clustering
  - Matrix factorization
- This includes a variety of important methods:
  - linear regression, logistic regression, non-parametric regression, nearest neighbours, decision trees, bagging, boosting, random forests, SVMs
  - PCA, K-means, matrix completion, topic modeling

Do I have the appropriate background?

Coursework is aimed at advanced undergrads and graduate students. We will use multivariate calculus, probability, and linear algebra.

- Linear algebra: vector/matrix operations such as eigenvalues and eigenvectors, eigen and singular value decompositions, inverse, trace, norms.
- Calculus: partial derivatives/gradient.
- Probability & Statistics: expectation, variance, covariance; Bayes' theorem; common distributions; maximum likelihood estimation, simple linear regression, point and interval estimation, hypothesis testing, p-values.

Do I have the appropriate background?

- Programming language: we are using R in this course.
  - Useful resources: `https://cran.r-project.org/`. A good review of some basic R commands is in Chapter 2.3 of the textbook.
  - How much do you need to know?
    - Basic knowledge on R is required (e.g., load data, create a vector or matrix, etc.)
    - The tutorials will provide you demonstrations of using R to perform statistical analysis.
    - The emphasis of coding will be on the use of the various R packages and on the implementation of the key subroutines of ML methods.
    - You will not be required to **implement complicated machine learning algorithms** nor to **write an entire R package**.

Textbook and other suggested readings

We mainly use the textbook

- Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*.
- You can find it via `https://www.statlearning.com`.

You are only responsible for the material covered in lectures, tutorials, and homeworks. You may also find the following references useful throughout the course:

- Hastie, Tibshirani, and Friedman. *The Elements of Statistical Learning*.
- Christopher Bishop. *Pattern Recognition and Machine Learning*.
- Kevin Murphy. *Machine Learning: a Probabilistic Perspective*.

There are lots of other freely available, high-quality ML resources.

Course Information

Course website: the main source of information; check regularly!
`http://courses.utstat.utoronto.ca/sta314_f23/`

Course email: sta314@utoronto.ca.

Crowdmark: homework releasing & submission, grades.

Quercus: annoucements only.

Piazza: the main place for discussions. Sign-up:
`https://piazza.com/utoronto.ca/fall2023/sta314h1`

Email policy

- Please, *do not* send emails to either the instructor or the TAs' personal / professional emails, except for absolutely urgent requests.

- For questions / requests,
  - if it is about requests such as extension of homework, regrading request, declaration of absence, missing the exam, etc., use the course email.
  - if it is related with solving homework problems, use the office hours.
  - if it is about course material such as lectures / tutorials, or clarification question, post it to Piazza so that other students will benefit.

Course delivery instructions

- All sections (LEC0101 and LEC0201) have the same delivery layout.

Weekly	delivery mode
2-hour lecture	in-person
1-hour tutorial	in-person
4-hour office hour (1 + 3)	zoom / in-person

- Tutorials are not required but highly recommended as they contain supplementary materials to the lectures. Some weeks might not have tutorials.
- While cell phones and other electronics are not prohibited in lecture, talking, recording or taking pictures in class is strictly prohibited without the consent of your instructor. Please ask before doing!

Course Information

All information is in the syllabus on the course website.
If you remember just one thing:

Check the course website regularly.

Grading

- (40%) 4 assignments
  - Combination of pen & paper derivations and light-weight programming exercises.
  - Weighted equally.
  - Hand-in on Crowdmark.
- (30%) Midterm test
  - 2-hour held during normal class time
  - See the syllabus for exact time and date.
  - Location to be announced (TBA).
- (30%) Final test
  - 2-hour held during the final assessment period
  - Date, time and location are TBA.

More on Assignments

Students are required to work on the assignments and submit their handouts alone. Discussion with instructors and other students is allowed. If you choose to do so, then you:

- Must include a statement in your submission that includes the name of the student that you discussed with and what part of your submission is involved.
- Must not share proofs, pseudocode, code, or simulation results.
- Must do your own work.

Violation of this policy is an academic offence and will be investigated and reported as such.

Assignments should be handed in by deadline; a late penalty of 10% of the total credit of the assignment per day will be assessed thereafter (up to 3 days, then submission is blocked).

Extensions will be granted only in special situations, and you will need to complete an absence declaration form and notify us to request special consideration, or otherwise have a written request approved by the course instructors at least three days before the due date.

Similar course: CSC311–Introduction to Machine Learning

- **STA314 takes a more statistical perspective than CSC311** while their core contents share the same machine learning methods.
  - The course will focus on the methodology and statistical insight rather than algorithm (or coding).
  - We do not cover neural networks nor reinforcement learning.
  - We will cover model selection, high-dimensional statistics, bootstrap, etc.

Advanced Courses

This course will help prepare you for the following courses.

- **STA414** (Statistical Methods for Machine Learning II)
  - This course is the follow-up course, which delves deeper into the probabilistic interpretation of machine learning.
- **CSC413** (Neural Networks and Deep Learning)
  - This course covers deep learning and automatic differentiation.
- **CSC412** (Probabilistic Learning and Reasoning)
  - The CSC analogue of STA414.

Questions on logistics?

What is learning?

"The activity or process of gaining knowledge or skill by studying, practicing, being taught, or experiencing something."

Merriam Webster dictionary

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

Tom Mitchell

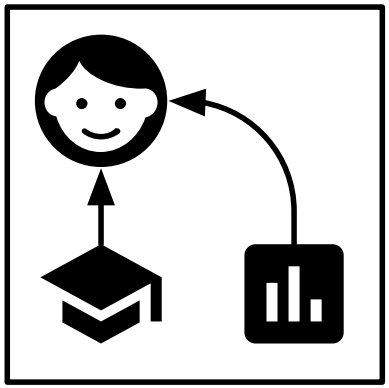
What is machine learning?

- Machine learning approach: program an algorithm to automatically learn from data, or from experience

- Why might you want to use a learning algorithm?
  - hard to code up a solution by hand (e.g. vision, speech)
  - system needs to adapt to a changing environment (e.g. spam detection)
  - want the system to perform *better* than the human programmers

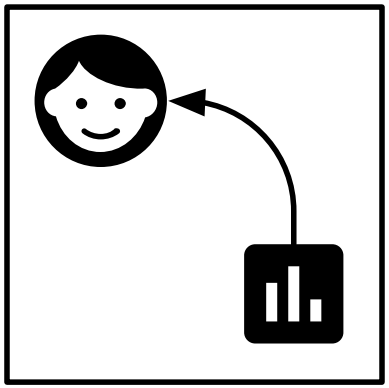
Types of machine learning problems

Supervised Learning



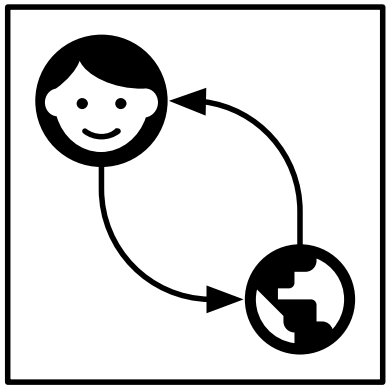
Machine is given data and examples of what to predict.

Unsupervised Learning



Machine is given data, but not what to predict.

Reinforcement Learning



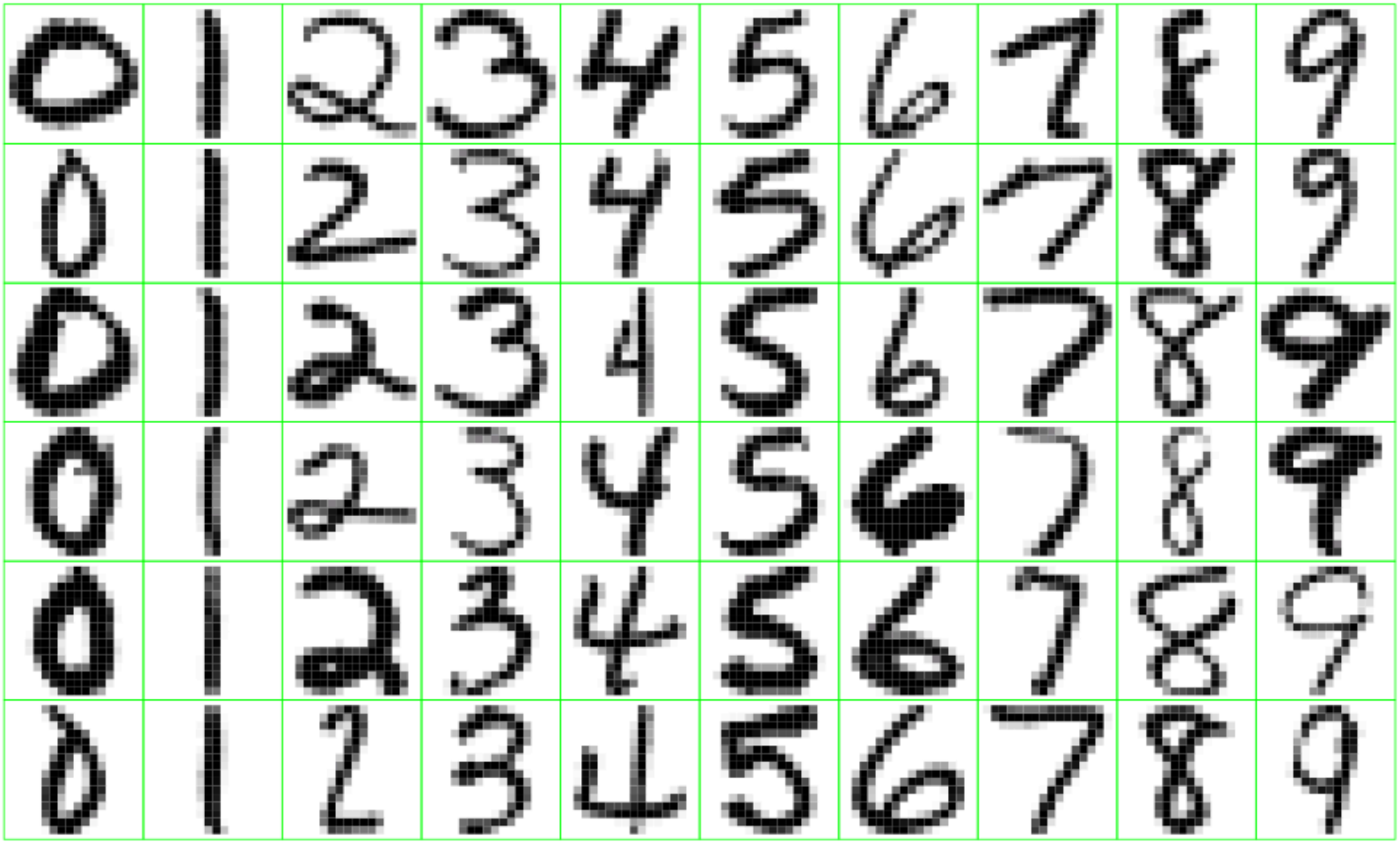
Machine gets data by interacting with an environment and tries to minimize a cost.

Spam emails detection

- data from 4601 emails sent to an individual (named George, at HP labs). Each is labeled as **spam** or **email**.
- goal: build a customized spam filter
- input features: relative frequency of 57 words and punctuation marks in the email message

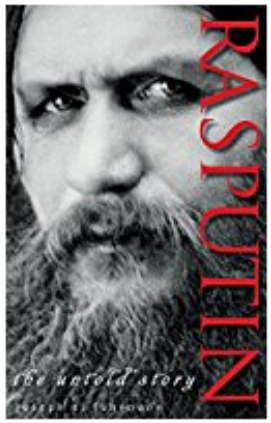
	george	you	hp	free	!	edu	remove
spam	0.00	2.26	0.02	0.54	0.51	0.01	0.28
email	1.27	1.27	0.90	0.07	0.11	0.29	0.01

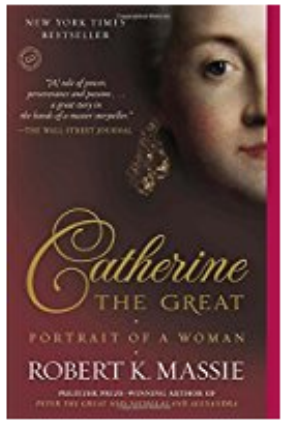
Detect numbers in a handwritten zip code

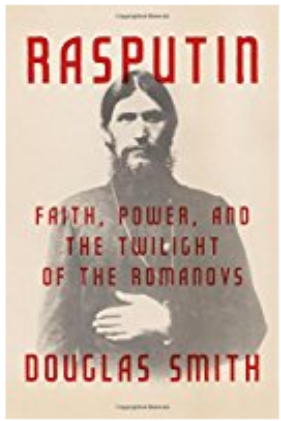


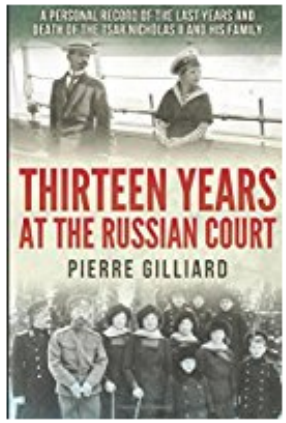
Recommender Systems : Amazon, Netflix, ...

Inspired by your shopping trends

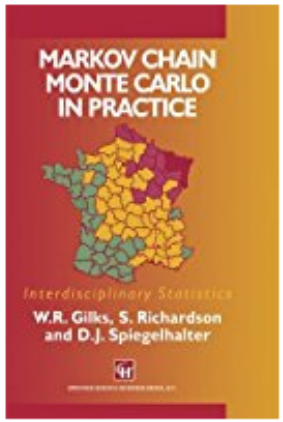


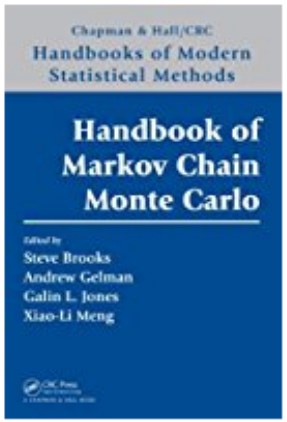


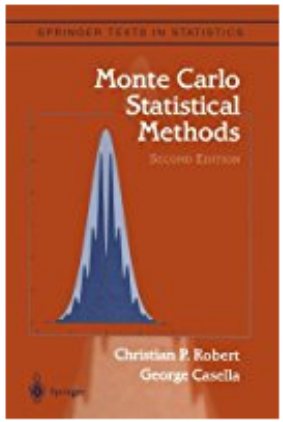


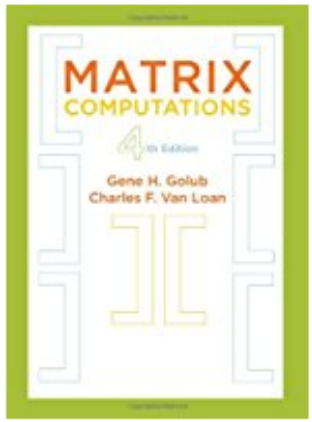


Related to items you've viewed See more









Computer vision: Object detection, semantic segmentation, pose estimation, and almost every other task is done with ML.

Object detection

DAQUAR 1553

What is there in front of the sofa?

Ground truth: table

IMG+BOW: table (0.74)

2-VIS+BLSTM: table (0.88)

LSTM: chair (0.47)

COCOQA 5078

How many leftover donuts is the red bicycle holding?

Ground truth: three

IMG+BOW: two (0.51)

2-VIS+BLSTM: three (0.27)

BOW: one (0.29)

Image segmentation

Speech: speech to text, personal assistants, speaker identification...

Natural language processing: machine translation, sentiment analysis, topic modelling.

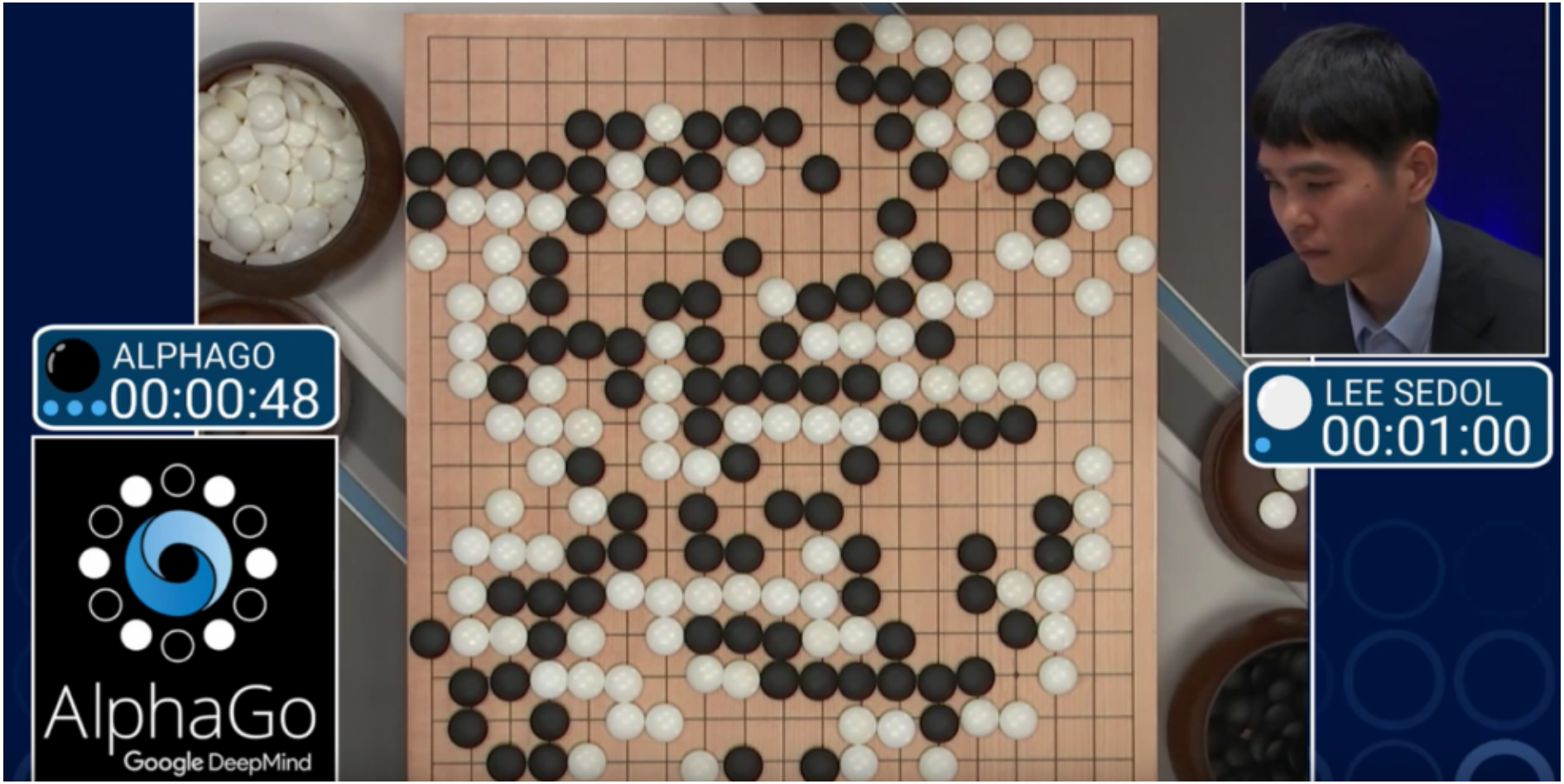
Real world example:

The New York Times

LDA analysis of 1.8M New York Times articles:

music band songs rock album jazz pop song singer night	book life novel story books man stories love children family	art museum show exhibition artist artists paintings painting century works	game knicks nets points team season play games night coach	show film television movie series says life man character know
theater play production show stage street broadway director musical directed	clinton bush campaign gore political republican dole presidential senator house	stock market percent fund investors funds companies stocks investment trading	restaurant sauce menu food dishes street dining dinner chicken served	budget tax governor county mayor billion taxes plan legislature fiscal

Playing Games



DOTA2 - ▸ Link

Statistical Learning versus Machine Learning

- Machine Learning (ML) is a subfield of Artificial Intelligence while Statistical Learning (SL) is a subfield of Statistics.
- They both try to uncover patterns in data.
- Both fields draw heavily on calculus, probability, and linear algebra, and share many of the same core algorithms
- ML puts more emphasis on algorithms, computation and prediction accuracy while SL emphasizes more on models and their interpretability, and how to evaluate uncertainty of the learning procedure.
- This course focuses on **Statistical Learning**.

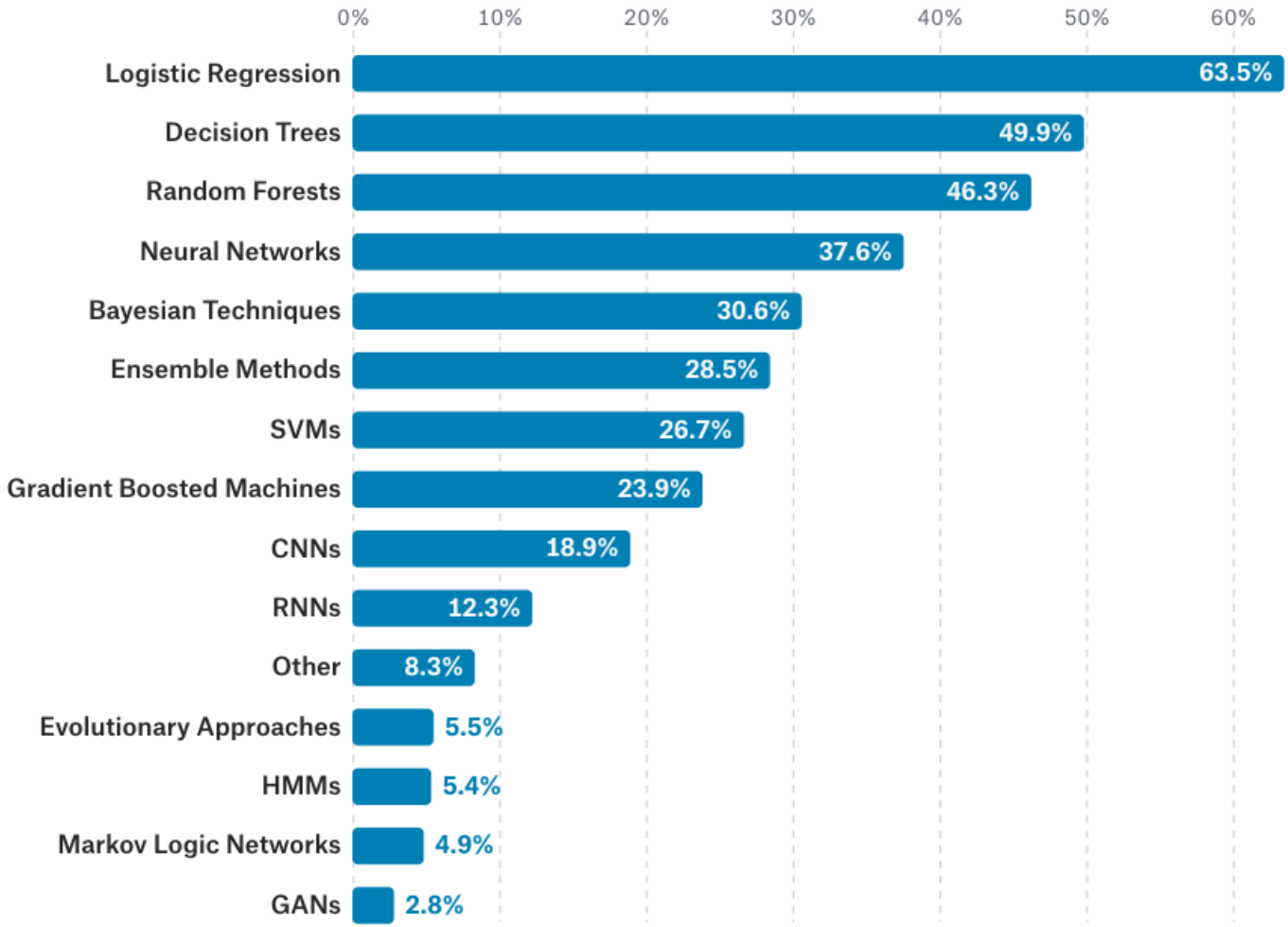
Why this class?

"I've heard that neural networks solve everything, can we just learn those?"

- There's a whole world of problems where neural nets do not work.
- The techniques in this course are still the first things to try for a new ML problem.
  - E.g., try logistic regression before building a deep neural net!
- The principles you learn in this course will be essential to understand and apply neural nets.

Why this class?

2017 Kaggle survey of data science and ML practitioners: what data science methods do you use at work?



Why this class?

- It is important to understand the ideas behind the various techniques, in order to know how and when to use them.
- Advanced algorithms are built on the simpler ones.
- It is important to accurately assess the performance of a method, to know how well or how badly it works or will work
- This is an exciting research area, having important applications in science, industry and finance.
- Statistical learning is a fundamental ingredient in the training of a modern data scientist.