

Machine Learning



Kevin Moon (kevin.moon@usu.edu)
STAT/CS 5810/6655





Outline

1. Course logistics
2. Introduction to Machine Learning
3. Course Overview
 1. Supervised Learning
 2. Unsupervised Learning
 3. Information Theory in Machine Learning
 4. Other potential topics
4. k -nearest neighbor classifier



Course Logistics

- Textbooks (recommended but not required; available online)
 - Introduction to Statistical Learning (ISL)
 - Elements of Statistical Learning (ESL)
- Many lecture notes are adapted from EECS 545 at Michigan
 - http://web.eecs.umich.edu/~cscott/past_courses/eecs545f16/index.html



Course Logistics



- Required background
 - Basic probability
 - Basic linear algebra & calculus
 - Programming experience
 - Look at HW 1 for an idea
 - See Syllabus for more specifics
- Canvas will be used for nearly everything (HW, announcements, etc.)
- Piazza will be used for answering most questions
 - An online forum dedicated to the class
 - Please reserve only private matters for email (use the features of Piazza for anonymity if desired)



Course Logistics

- Office hours: Tues 2:30-4:30, Thurs 1:30-3:15
- TA: Haozhe Chen
- TA Office hours: TBD
- 6 HW assignments (60% of grade)
 - Assigned about every 2 weeks, due on Fridays
 - All/most will include some programming
 - Collaboration is allowed but you MUST indicate whom you worked with
 - Any sources you used (e.g. Stack exchange, ChatGPT) must be cited as well
 - **5810:** Lowest homework score will be dropped and all homeworks will be less mathy
 - See Syllabus for details on late homework policies



Course Logistics

Final project (40% of grade)

- In groups of 2-3
- Each group must select a real problem and data
 - Venture beyond MNIST and UCI repository
 - Kaggle could be a good source
- Apply some machine learning (ML) methods to the problem
 - Exact scope will differ between the classes (5810 vs 6655)
- Proposal due around the end of February
- A progress report and final report will also be required
- Final presentation may be required of 6655



Expected workload



- 3 credit graduate level course ⇒ minimum of 9 hours per week outside of class
 - Most students will probably spend more
- This is an advanced graduate course
 - I expect you to do some learning on your own
 - I won't always give you exactly everything you need to complete every problem. You may need to consult resources provided on Canvas as well as other resources
- That said, ask questions!
 - In class
 - In office hours
 - On Piazza



Some thoughts on proofs

- I highly recommend you start with scratch paper (or a whiteboard) when trying to prove or show something
- Write down your official answer after you have figured out all of the steps
- This will improve the legibility of your answers
 - Less erasing or crossing out needed
- It is also mentally liberating
 - You no longer have to worry about needing to erase things if you take a wrong direction





A little about me...

- Grew up in Duchesne, Utah
- B.S. in electrical engineering from BYU (2011)
 - Minors in math and economics
- M.S. in electrical engineering from BYU (2012)
 - Radar image analysis
- M.S. in mathematics from the University of Michigan (2016)
- Ph.D. in electrical engineering from Michigan (2016)
 - Machine learning, data science, information theory, etc.
- Postdoc in the Genetics department and the Applied Math program at Yale University





Goals of the Course



- Gain knowledge of and familiarity with selected classical and advanced machine learning techniques and principles that will be useful in both the university and industry settings
- Learn to apply the selected techniques to solve problems with real data
- Learn selected general principles (e.g. optimization, the kernel trick) for understanding machine learning methods not covered in the course



Why theory???



- Machine learning is math
- Programming is math
- Proofs are a way of solving problems in math
- The better you are at the theory (e.g. proofs), the better your problem-solving skills will be
 - This will bleed into solving problems that aren't directly proofs
- Machine learning is starting to be automated...
 - E.g. AutoML by Google
 - The more you understand about the theoretical principles of ML, the less likely you'll be replaced by automated data scientists
- After this course, you should be able to read and understand most ML research papers (because of the math)



Student testimonials on theory

- “The theoretical portions of the homework were challenging and at times fairly time consuming. However, I can honestly say that at the end of the semester, I was glad that I had stuck it out. The understanding that I gained about the mechanics of the various methods and algorithms has increased not only my proficiency in machine learning methods, but also improved my ability to break down a problem, understand the mathematics of the problem, and solve the problems in a systematic way. This class did stretch me, but as a result I felt that this class helped me grow in a variety of ways that has improved my performance in my research and in other courses.” –an M.S. STAT student
- “Even with my long industry experience as a Data Scientist, I found this Machine Learning course as incredibly valuable. I thought I already knew a lot about various ML Algorithms but after completing the course I now feel that I lacked depth in my prior understanding. The lectures were stimulating and the homework assignments really helped build my understanding of the Math behind Machine Learning. I have gained significant confidence in reading Journal Papers and I have also noticed that I am able to have much deeper conversations with my peers on ML topics, all because of depth of understanding and confidence gained through this course. This is a very high quality course and matches in rigor and intensity of ML courses taught at leading universities in the country.” –Ronak Tali
- “Previous stat classes I’d taken leaned heavily toward application, so initially I was a bit resistant to the focus on theory. ‘In data science and analytics jobs aren’t we paid for the models we develop and the code that creates them?’ was my main critique. However, in talking to industry practitioners and browsing online forums like Quora and Reddit, I came to realize it’s not that simple. A huge benefit of understanding the theory behind ML is knowing when to apply a technique. In many real-world cases, the prediction may be of secondary importance. For example, model interpretation or statistical confidence may be important. Understanding the theory can help with these and other questions.” –another M.S. STAT student



Wait, I've also heard about statistical learning...

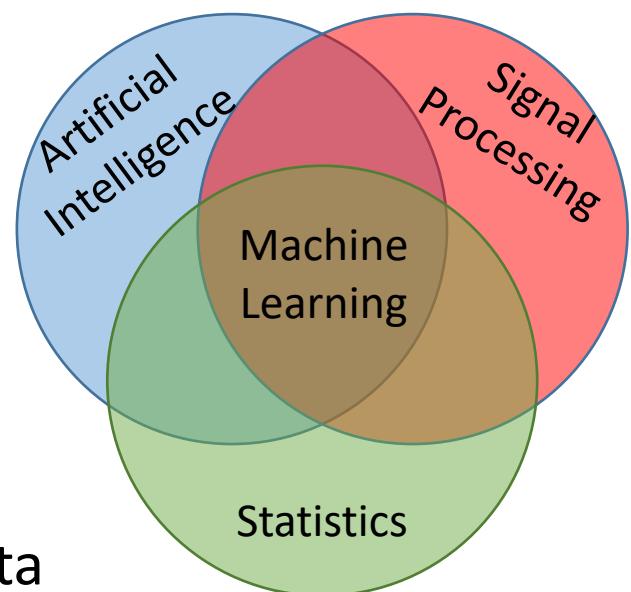


- Statistical Learning = Machine Learning (to me)
 - “Statistical” learning emphasizes modern ML’s roots and foundation in probability and statistics for handling uncertainty
- So what is statistical/machine learning?
 - A field of study concerned with making quantitative inferences and predictions from data
 - Uses statistical techniques and computational power to “learn” these inferences and predictions without being explicitly programmed
 - A little more formally: an algorithm is said to learn from data if its performance improves as the amount of data increases



Roots of machine learning

- ML theory and methodology emerged from three main areas: artificial intelligence, signal processing, and statistics
- Examples of contributions:
 - AI: Support vector machines, neural networks
 - SP: information theory, k-means, detection and estimation theory
 - Stats: regression, classification, a probabilistic framework for handling uncertainty and randomness in the data





Classical Statistics vs ML

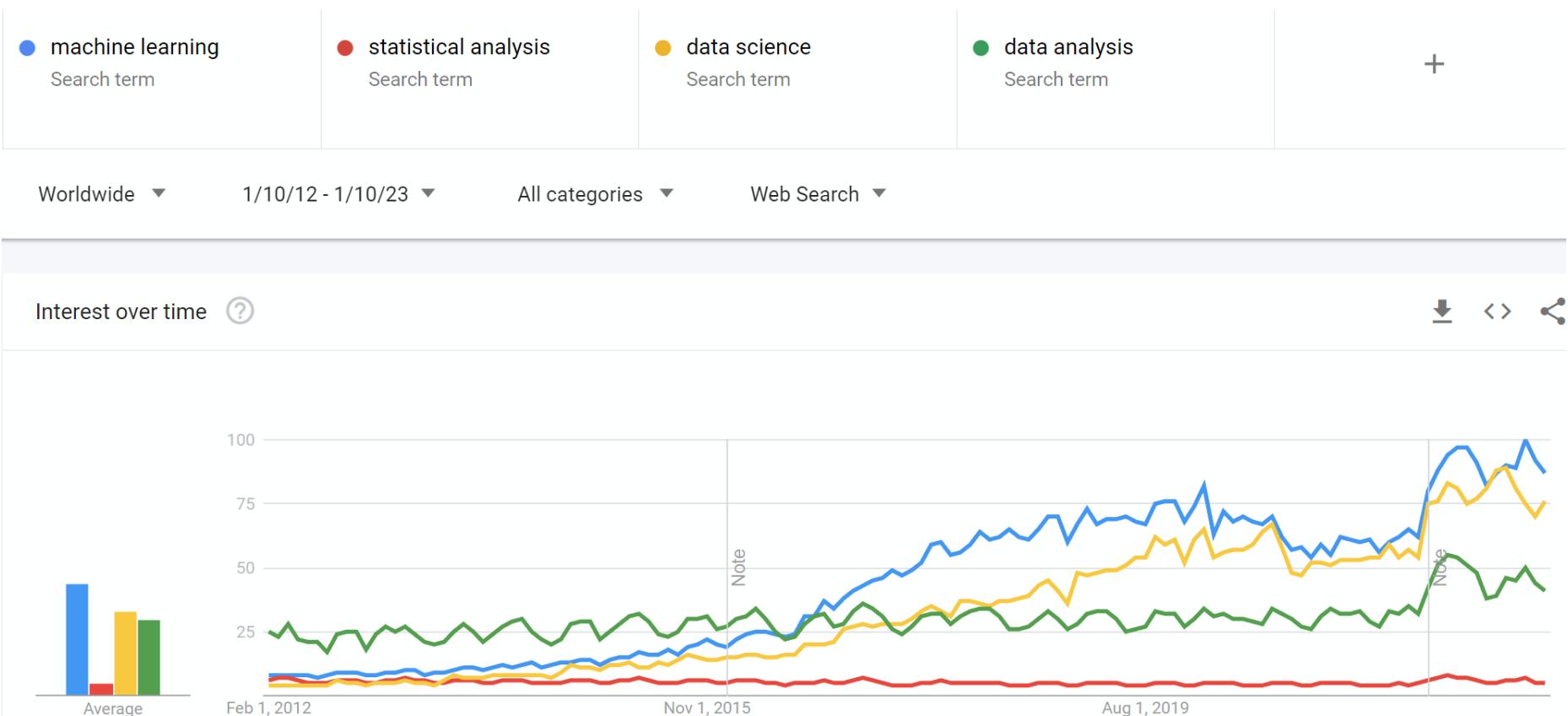


- What are the differences between classical statistics and ML?
 - There is an overlap of methods
 - Somewhat broadly, classical statistics focuses heavily on data modeling
 - ML typically focuses on the endgame/output (e.g., prediction)
 - A model is useless if it doesn't give a good prediction
 - Many ML methods are based on statistical models
 - See Breiman (2001) for one viewpoint on the two fields
 - **Assigned for HW 1**



Why study machine learning?

- It certainly is popular...





Machine Learning is hot



- NeurIPS 2018 sold out in less than 15 minutes

Ben Hamner ✅ @benhamner

Looks like NIPS 2018 may have sold out in under 15 minutes. For those debating ML hype, getting a ticket to a ML conference is now more challenging than a Taylor Swift conference or a Hamilton showing

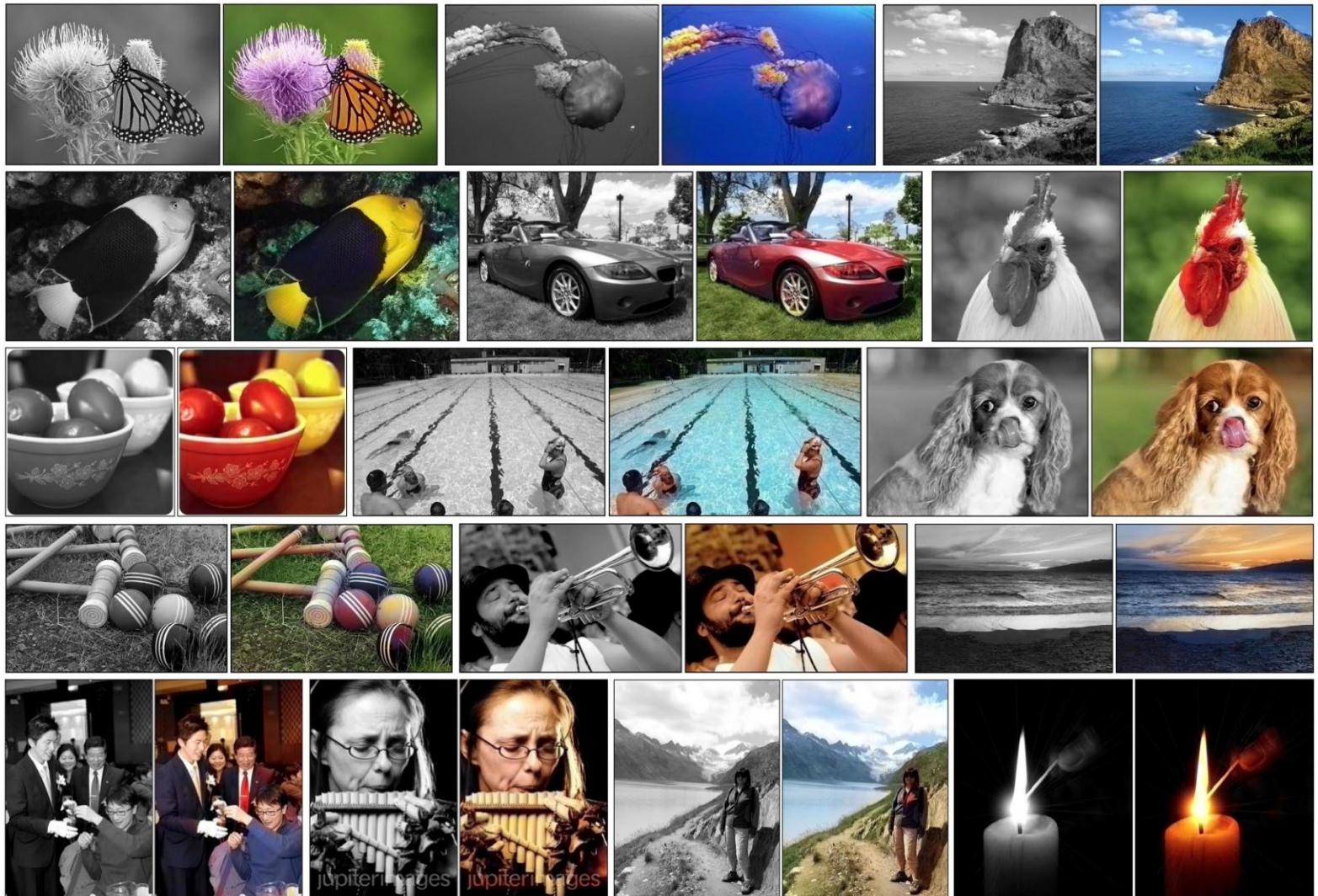
[Show this thread](#)

9/4/18, 6:56 PM



Success of Machine Learning

Image colorization (Zhang et al., 2016)





Success of Machine Learning

Image colorization (Zhang et al., 2016)



Colorized classical photographs by Ansel Adams



Success of Machine Learning



Real-time visual translation on smartphones

1. Find the letters
2. Recognize the letters
3. Translate
4. Render the translation in the same style



MÖRK
Mörk → Dark

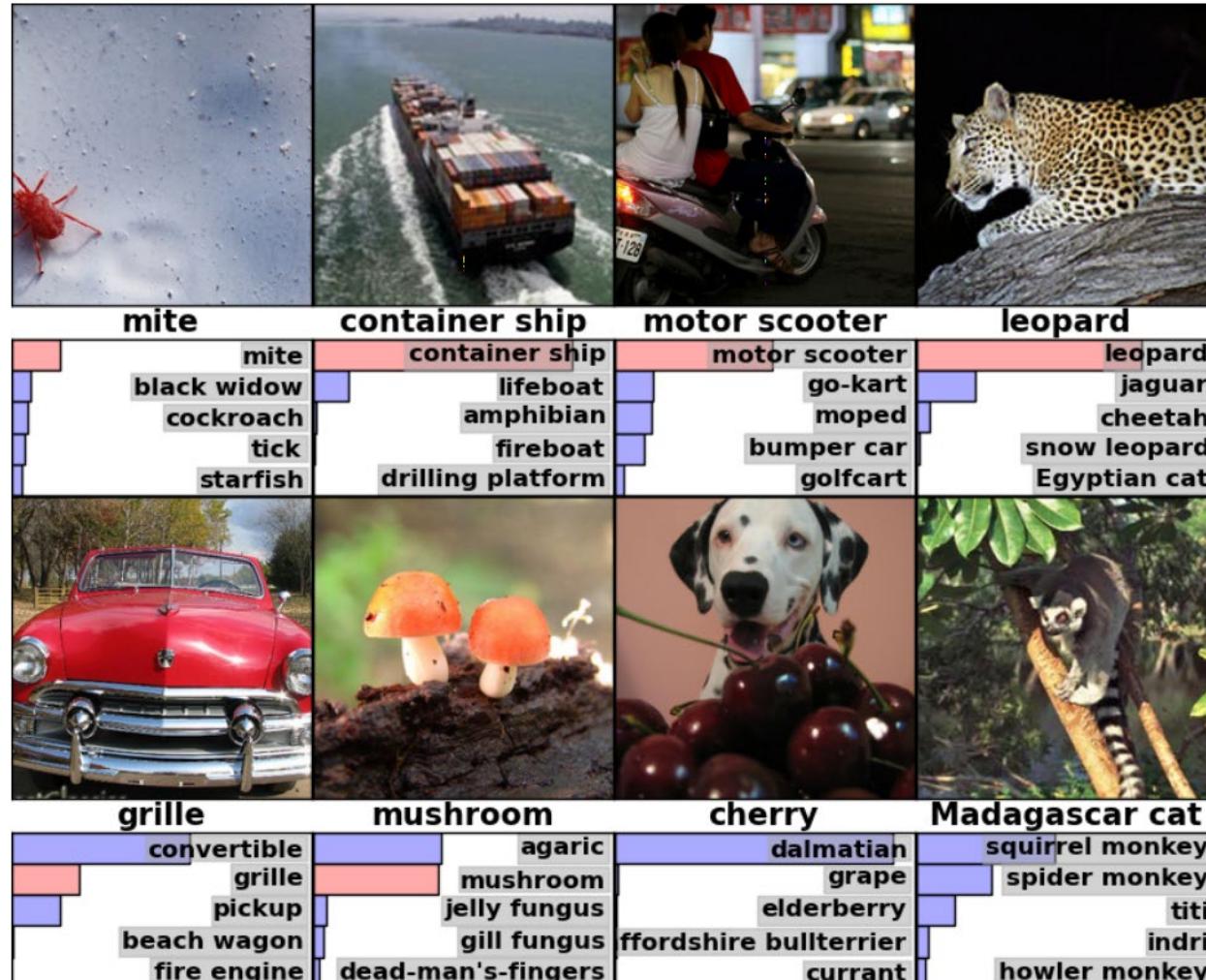


Google blog, 2015



Success of Machine Learning

Object classification/detection in images (Krizhevsky et al., 2012)





Success of Machine Learning



Automatic image caption generation (Karpathy & Fei-Fei, 2015)



"man in black shirt is playing guitar."



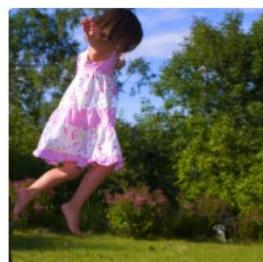
"construction worker in orange safety vest is working on road."



"two young girls are playing with lego toy."



"boy is doing backflip on wakeboard."



"girl in pink dress is jumping in air."



"black and white dog jumps over bar."



"young girl in pink shirt is swinging on swing."



"man in blue wetsuit is surfing on wave."



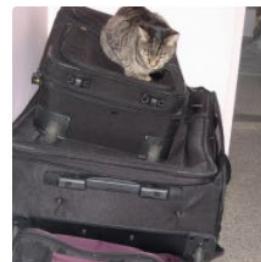
"little girl is eating piece of cake."



"baseball player is throwing ball in game."



"woman is holding bunch of bananas."



"black cat is sitting on top of suitcase."



Success of Machine Learning

Automated essay writing with ChatGPT

Reset Thread

Dark Mode

OpenAI Discord

Learn More

Log out

C Can you write me a 4 paragraph academic essay comparing and contrasting the theories of nationalism of Benedict Anderson and Ernest Gellner?

G Benedict Anderson and Ernest Gellner both have theories about the origins and nature of nationalism, but their perspectives differ in key ways. Anderson believes that nationalism is a cultural construct, arising from the development of print capitalism and the creation of imagined communities. Gellner, on the other hand, sees nationalism as a product of modernization and industrialization, with nations being formed as a means of organizing society and providing a shared identity.

One key difference between the two theories is the role of culture. For Anderson, culture is central to the creation of nationalism, as the spread of print media allows for the creation of imagined communities that transcend physical boundaries. This allows for the development of a national identity based on shared cultural experiences and practices. In contrast, Gellner downplays the importance of culture in the formation of nationalism, arguing that it is primarily a function of economic and political developments. He sees nations as being formed in response to the need for a shared identity and sense of belonging in an industrialized society.

Another point of contrast is the emphasis on agency. Anderson's theory suggests that people actively participate in the construction of national identity, through their engagement with print media and other cultural practices. In contrast, Gellner's theory emphasizes the role of external factors in shaping national identity, with individuals being passive recipients of the national identity imposed on them by the forces of modernization.

Overall, while both Anderson and Gellner provide valuable insights into the origins of nationalism, their theories offer different perspectives on the role of culture and agency in the formation of national identity. Anderson emphasizes the cultural construction of nationalism, while Gellner sees it as a product of modernization and industrialization. Both theories offer valuable insights into the complex process of national identity formation, and can be useful in understanding the dynamics of nationalism in different historical contexts.

< 3 / 3 >

Up Down



Success of Machine Learning

Automated programming with ChatGPT



Find the bug with this code:

```
...
for (var i = 0; i < 5; i++) {
    setTimeout(() => console.log(i), 1000)
}
...
```



The code contains a bug that will cause it to always print the number 5 to the console, rather than the expected sequence of numbers from 0 to 4. This is because the `setTimeout` function creates a closure that captures the value of the `i` variable, but the `i` variable is changed before the callback function is executed.



To fix this bug, the code can be modified to create a new variable in the loop that is passed to the `setTimeout` callback function, rather than using the loop variable directly. This will ensure that each callback function captures the correct value of the loop variable, and the expected output will be produced. Here is an example of how the code could be fixed:

Copy code

```
for (var i = 0; i < 5; i++) {
    // Create a new variable and pass its value to setTimeout
    const value = i;
    setTimeout(() => console.log(value), 1000)
}
```

This code will print the numbers 0 to 4 to the console, as expected. Alternatively, the `let` keyword can be used in place of `var` to declare the loop variable, which will automatically create a new variable for each iteration of the loop and avoid the need to create a new variable manually. This is a common pattern for avoiding closure-related bugs in JavaScript.



Success of Machine Learning

AI generated art



Best Dalle2 Pics
@Dalle2Pics

...

A raccoon playing tennis at Wimbledon in the 1990s
[#dalle2 #dalle](#)



3:33 AM · May 14, 2022



Success of Machine Learning

Automatic game playing

Alpha Go Zero

nature
International Journal of science

Access provided by Yale University

Altmetric: 2188 Citations: 1 [More detail >](#)

Article

Mastering the game of Go without human knowledge

David Silver Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel & Demis Hassabis

Nature 550, 354–359 (19 October 2017)
doi:10.1038/nature24270
[Download Citation](#)
[Computational science](#) [Computer science](#) [Reward](#)

Received: 07 April 2017
Accepted: 13 September 2017
Published online: 18 October 2017

arXiv.org > cs > arXiv:1712.01815

Search or
(Help | Advanced)

Computer Science > Artificial Intelligence

Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, Demis Hassabis

(Submitted on 5 Dec 2017)

The game of chess is the most widely-studied domain in the history of artificial intelligence. The strongest programs are based on a combination of sophisticated search techniques, domain-specific adaptations, and handcrafted evaluation functions that have been refined by human experts over several decades. In contrast, the AlphaGo Zero program recently achieved superhuman performance in the game of Go, by tabula rasa reinforcement learning from games of self-play. In this paper, we generalise this approach into a single AlphaZero algorithm that can achieve, tabula rasa, superhuman performance in many challenging domains. Starting from random play, and given no domain knowledge except the game rules, AlphaZero achieved within 24 hours a superhuman level of play in the games of chess and shogi (Japanese chess) as well as Go, and convincingly defeated a world-champion program in each case.

Subjects: Artificial Intelligence (cs.AI); Learning (cs.LG)
Cite as: [arXiv:1712.01815 \[cs.AI\]](#)
(or [arXiv:1712.01815v1 \[cs.AI\]](#) for this version)

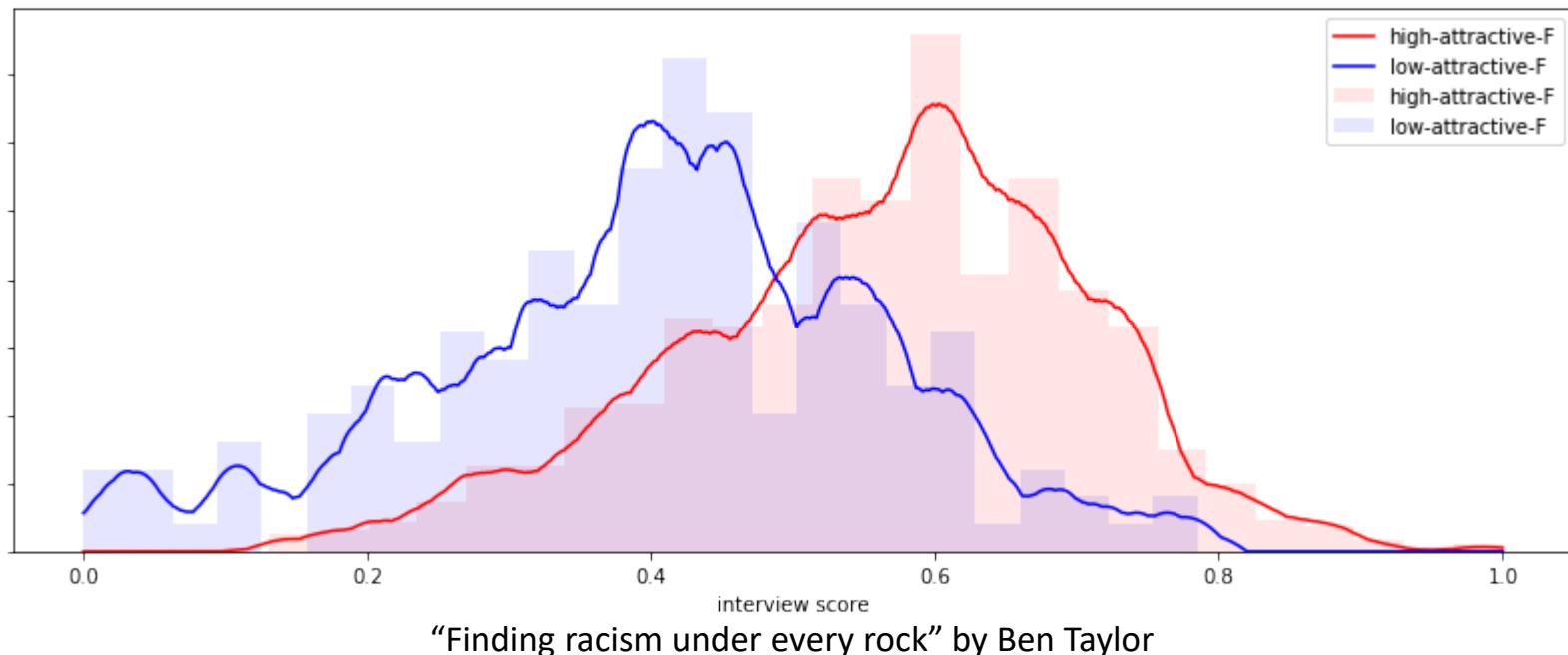
Submission history

From: David Silver [[view email](#)]
[v1] Tue, 5 Dec 2017 18:45:38 GMT (272kb,D)



Success of Machine Learning

- Finding bias in human-based job interviews
- More attractive women are rated higher by humans

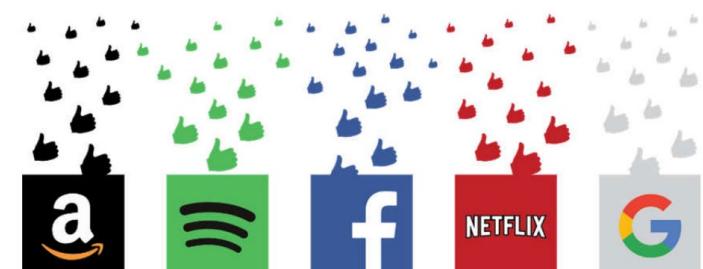


- Similar method finds evidence of some racial bias



Other Machine Learning Applications

- Chatbots (NLP)
- Fraud detection
- Route planning
- Search engine refinement
- Face recognition/detection
- Credit risk assessment
- Financial market prediction
- Medical diagnosis
- Personalized medicine
- Electricity demand forecasting
- Spam filtering
- Collision avoidance systems
- Speech synthesis/analysis (e.g. Siri, Alexa)
- Recommender systems (e.g. Netflix, Spotify)





Some notation and terminology

- Typically will denote a measurement as $\mathbf{x} \in \mathbb{R}^d$
- \mathbf{x} is usually a Euclidean vector, denoted

$$\mathbf{x} = \begin{bmatrix} x^{(1)} \\ \vdots \\ x^{(d)} \end{bmatrix} \quad \text{or} \quad \mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$$

- \mathbf{x} is a pattern, signal, input, instance, or feature vector
- x_i is a feature, attribute, predictor, or covariate
- \mathbf{x} is typically viewed as a realization of a random variable/vector \mathbf{X}



Supervised learning



- The learner/user are given training data

$$(x_1, y_1), \dots, (x_n, y_n)$$

- Each y_i is the output of an unknown and possibly noisy function with input \mathbf{x}_i
- Goal: Given a test input \mathbf{x} , predict the correct output
- Two primary types of supervised learning: classification and regression



Classification

- Outputs are called labels and are typically finite:

$$y \in \{1, 2, \dots, C\}$$

where C is the number of classes

- Example: handwritten digit recognition

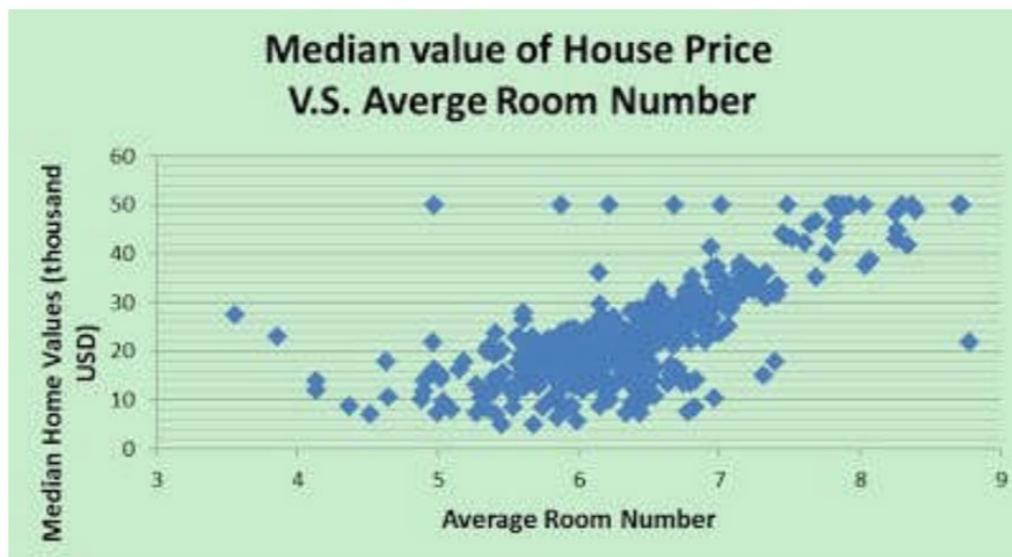




Regression



Outputs are called responses and are real valued: $y \in \mathbb{R}$





Unsupervised Learning

- Data is unlabeled
 $\mathbf{x}_1, \dots, \mathbf{x}_n$
- No outputs are given
- Goal: infer a property from the data, not to predict a property from future data
- Examples:
 - **Clustering:** What are the distinct clusters or groups in a dataset?
 - **Density Estimation:** What is the probability density function that generated the data?
 - **Dimensionality Reduction:** How can we reduce the dimensionality of the data without losing (much) information?



Other Machine Learning Problems



- Semi-supervised learning
- Online learning
- Reinforcement learning
- Active learning
- Anomaly detection
- Transfer learning
- Multi-task learning
- Matrix completion
- Learning to rank
- ...



Challenges in Machine Learning



- Big data (both sample size and dimension)
- Noisy labels
- Missing labels
- Missing features
- Data representations
- Safety (e.g. self-driving cars)
- Fairness/ethics
- Changing environment



This Course

- Exposure to important algorithms in ML
- Derivation of many of these algorithms from first principles
- Discussion of general phenomena that pervade machine learning (generalization, overfitting, regularization, etc.)



What about data mining?

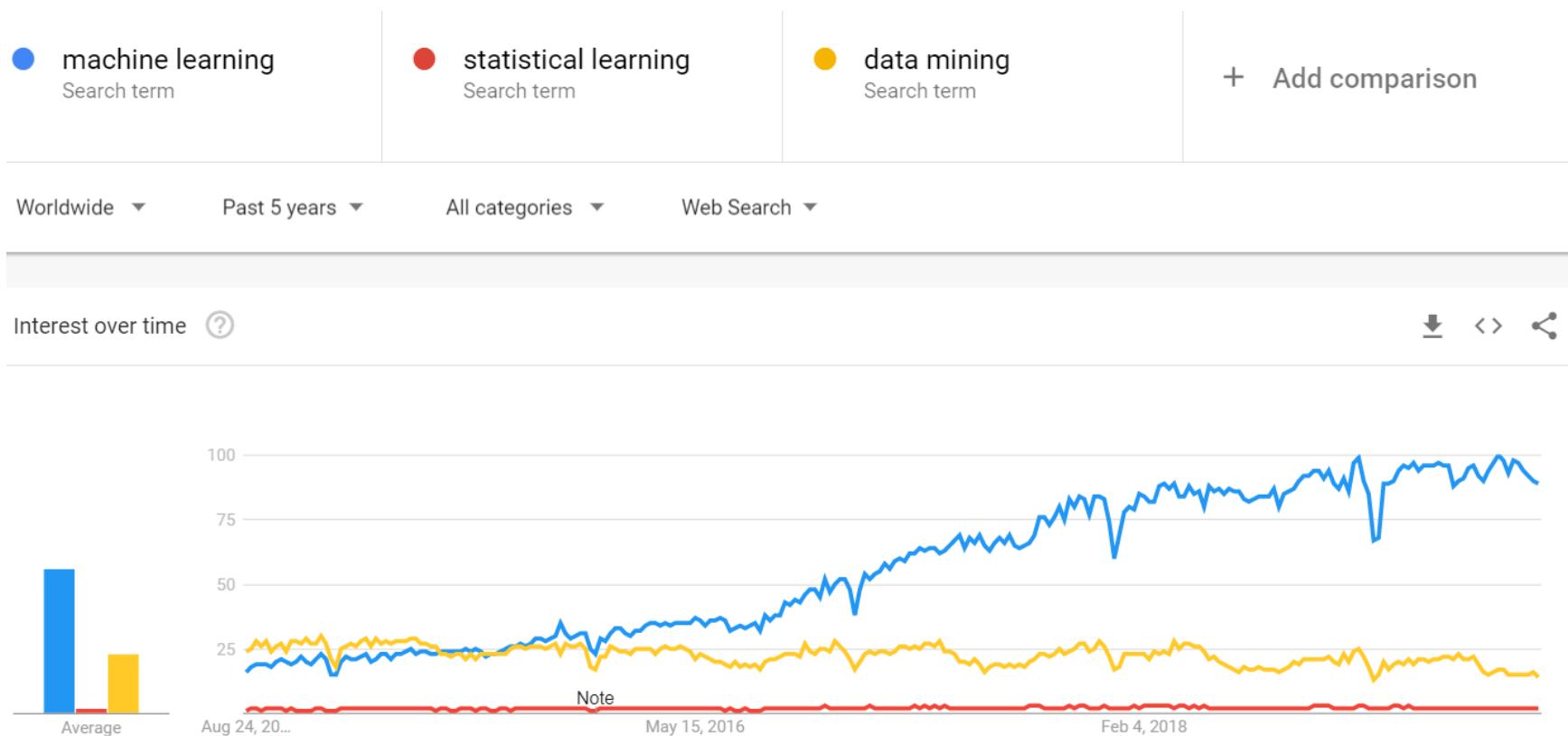
- What is data mining?
 - Depends on who you talk to...
- Data mining includes many topics in both supervised and unsupervised learning
 - Wikipedia claims data mining includes anomaly detection, dependency modeling, clustering, classification, regression, and summarization (e.g. visualization)
 - Also can include issues like database management (which we will not cover at all)
 - Some people say that data mining is ML applied to large datasets, but not everyone (including me) makes this distinction



What about data mining?



- In practice, machine learning has come to encompass essentially all aspects of data mining



- You probably won't hear the terms “data mining” or “statistical learning” again this semester



What is artificial intelligence?

- **Wikipedia:** "...a field of study in computer science that develops and studies intelligent machines." Breaks it down into sub-problems such as problem solving, knowledge representation, planning and decision making, natural language processing, perception, social intelligence, and general intelligence.
- **AWS:** "...the field of computer science dedicated to solving cognitive problems commonly associated with human intelligence, such as learning, creation, and image recognition."
- **IBM:** "...a field which combines computer science and robust datasets to enable problem-solving."



Artificial intelligence vs. machine learning

- As a field, AI generally includes machine learning but also includes more
- Most recent, popular “AI” advances are better classified as machine learning advances
- AI has become the latest buzzword, so many machine learning methods and problems are being rebranded as AI



Some adjectives used to describe ML algorithms

- **Distributional assumptions**

- Generative: assumes a full probabilistic model of the observed data
- Discriminative: assumes a partial or no probabilistic model

- **Computational form**

- Linear: produces a linear/affine function
- Nonlinear: not linear

- **Complexity**

- Parametric: # of model parameters* is independent of sample size
- Nonparametric: # of model parameters* grows with sample size

*Model parameters = parameters needed to represent the function produced by the learning algorithm

Course Overview



Topics in this Class

Supervised Learning

- Classification
 - Linear classifiers: LDA, logistic regression, separating hyperplane
 - Nonlinear classifiers: k-nearest neighbors, naïve Bayes, SVM, decision trees
- Regression: linear regression, regularized and kernel ridge regression
- Ensemble methods: boosting and bagging, random forests
 - Used for both regression and classification
- General principles
 - Empirical risk minimization and regularization, unconstrained and constrained optimization, the kernel trick, model selection
- Neural networks (a little)



Topics in this Class (cont.)



Unsupervised Learning

- Kernel density estimation
- Dimensionality reduction: PCA, Diffusion Maps, and PHATE
- Clustering algorithms: k-means, GMM, spectral clustering
- The EM algorithm
- Random projections/random PCA (if time)
- Matrix completion/imputation (if time)
- Information theory in machine learning
 - Entropy, divergence, and mutual information
- Reinforcement learning (if time)
- Other topics (if time)



Empirical Risk Minimization

- A general framework for supervised learning problems
- Define the problem in terms of a cost or risk function that should be minimized
- Different machine learning methods come from assuming different cost functions or applying different assumptions on the function that maps from inputs to outputs
- Regularization
 - Generally an addition to the cost function that imposes additional structural assumptions
 - E.g. sparsity, small norm



Optimization

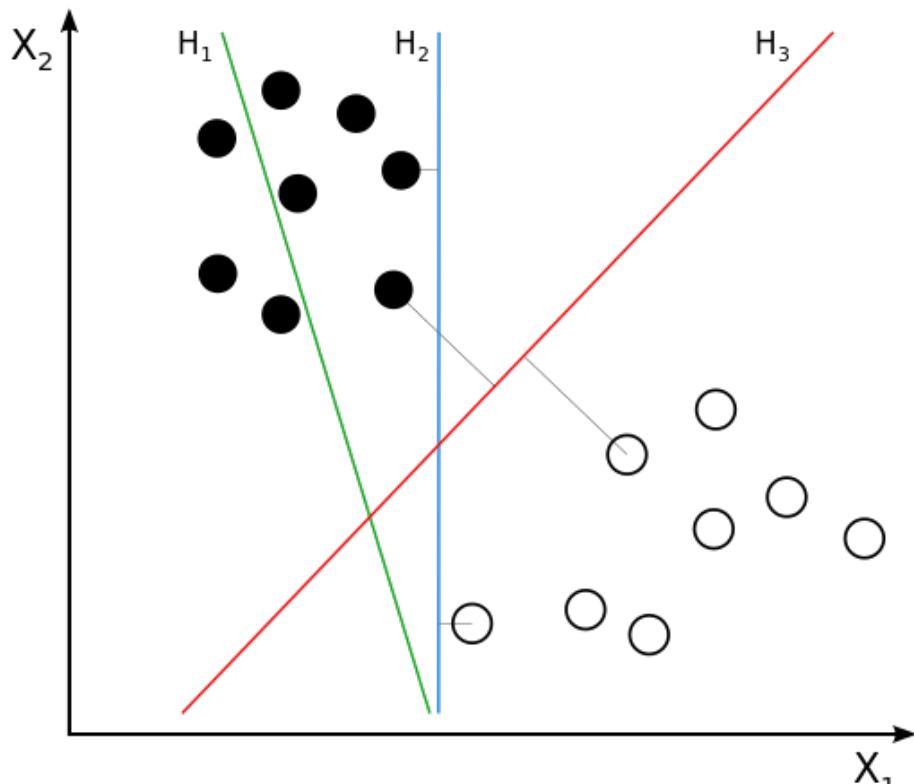
- Most cost functions in ML do not have a closed-form minimizer
 - This includes neural networks
- Optimization techniques allow us to find solutions
 - Global solution if the problem is “convex”
 - Local solution (may still be global) if the problem is not
- We'll cover constrained and unconstrained optimization



Linear classifiers



- Goal is to find a line (or hyperplane in higher dimensions) that separates classes
- Includes **LDA, logistic regression, and separating hyperplane**
- Separating hyperplanes use optimization theory to find a hyperplane that maximizes the margin between two classes

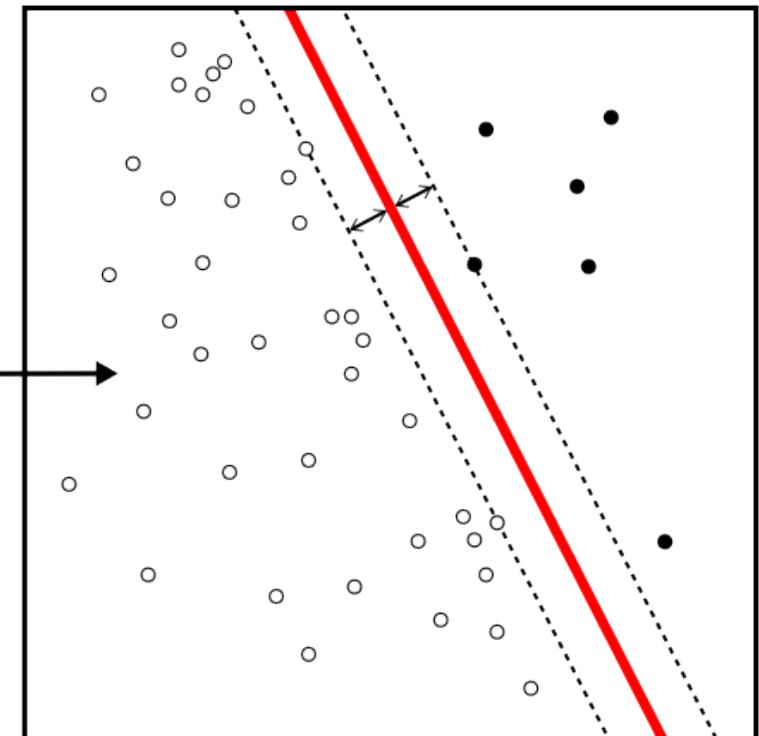
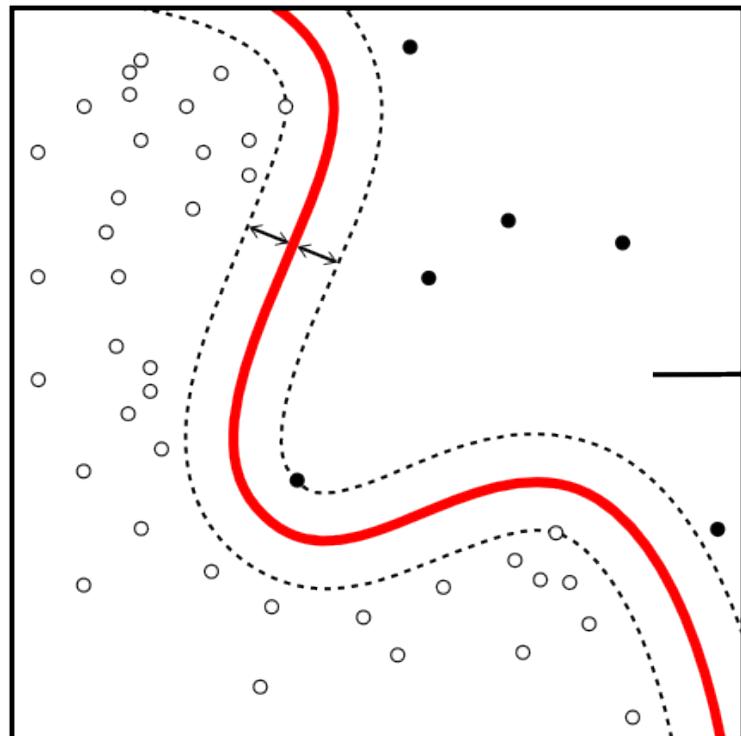


[https://commons.wikimedia.org/
w/index.php?curid=22877598](https://commons.wikimedia.org/w/index.php?curid=22877598)



The kernel trick and RKHS

- The kernel trick allows us to convert nonlinear problems into linear problems
- The theory of reproducing kernel Hilbert spaces (RKHS) shows we can even do this for infinite dimensional data



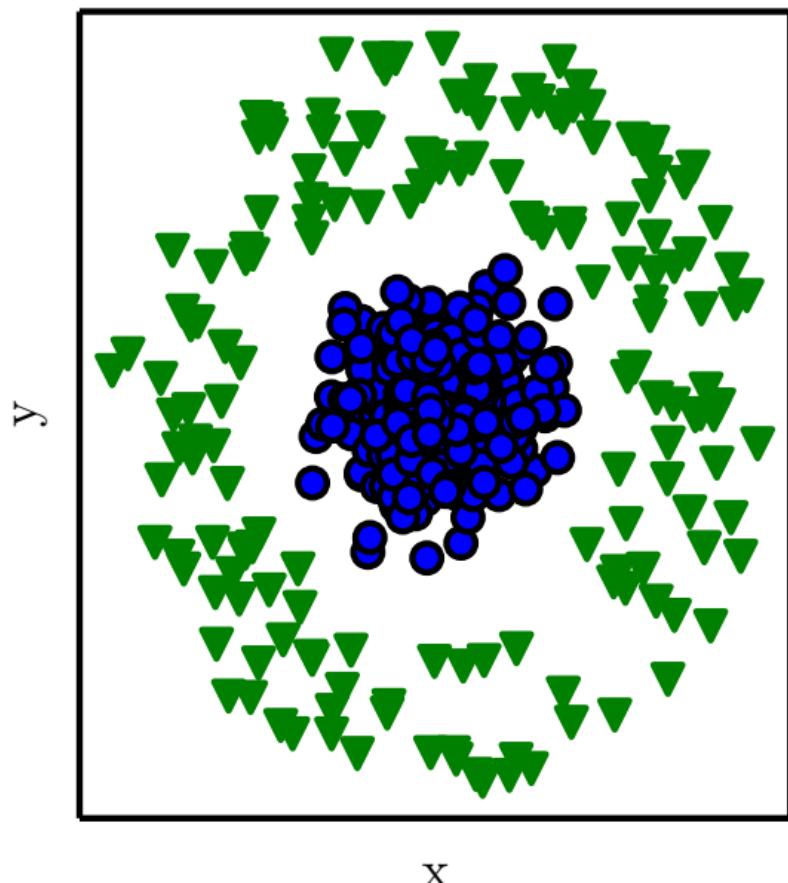
<https://commons.wikimedia.org/w/index.php?curid=14941564>



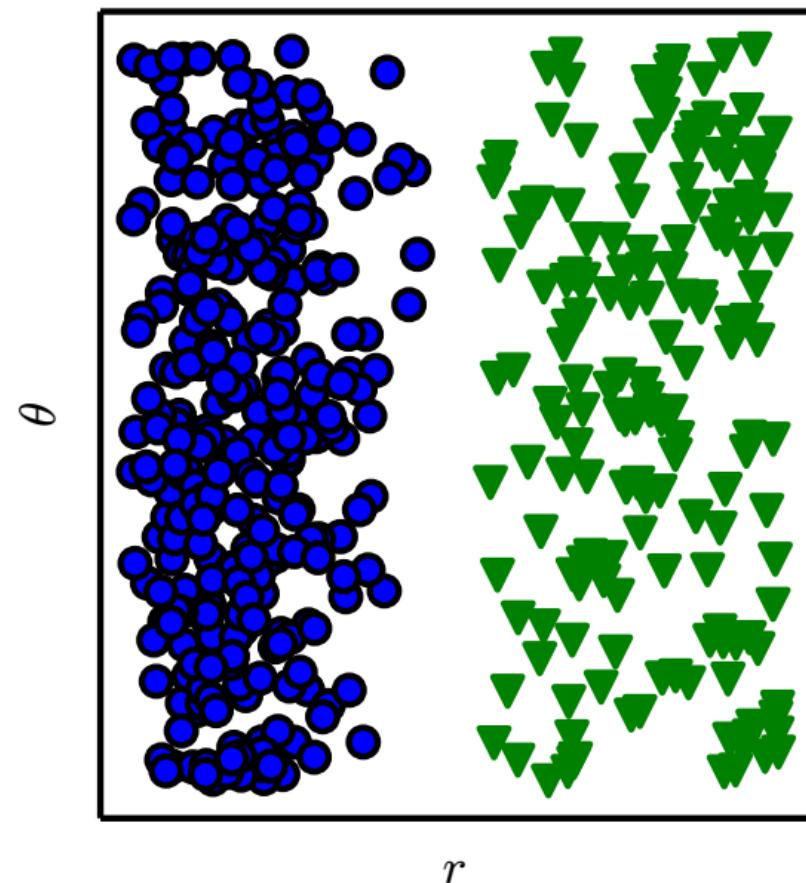
Ex. of different representation



Cartesian coordinates

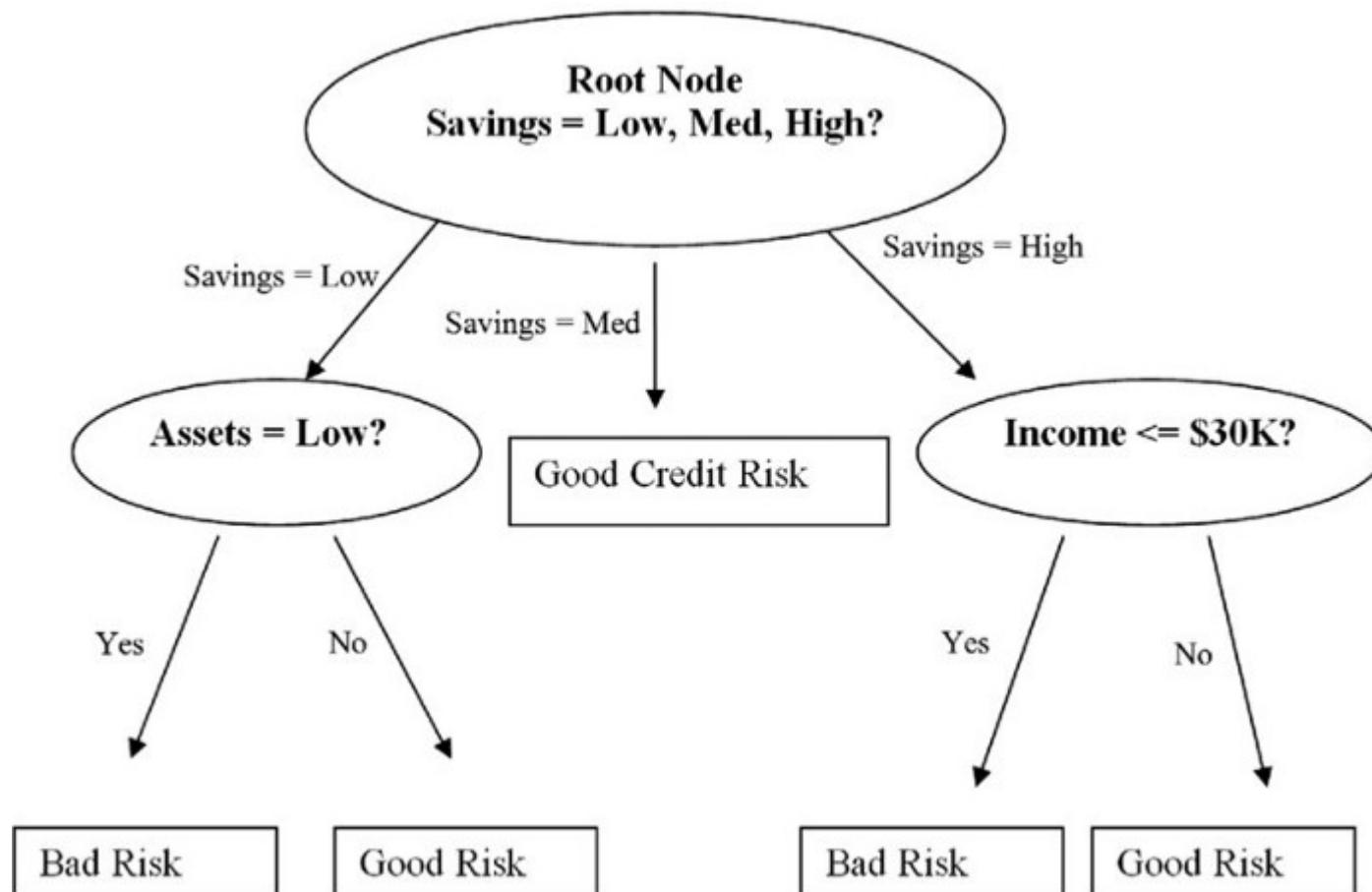


Polar coordinates



Goodfellow et al., 2016

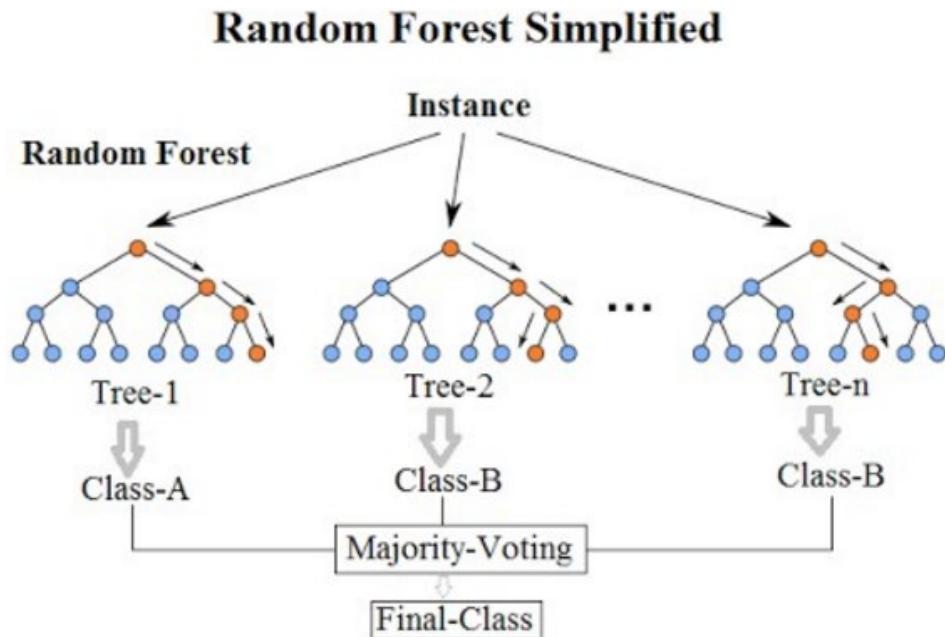
Decision trees





Ensemble methods

- Many simple ML methods are very flawed on their own
 - E.g., decision trees can perform poorly and are very sensitive to the training data
- Ensemble methods take a collection (ensemble) of weak (poor) learners and combine them to create a strong (good) learner
- E.g., **random forests**



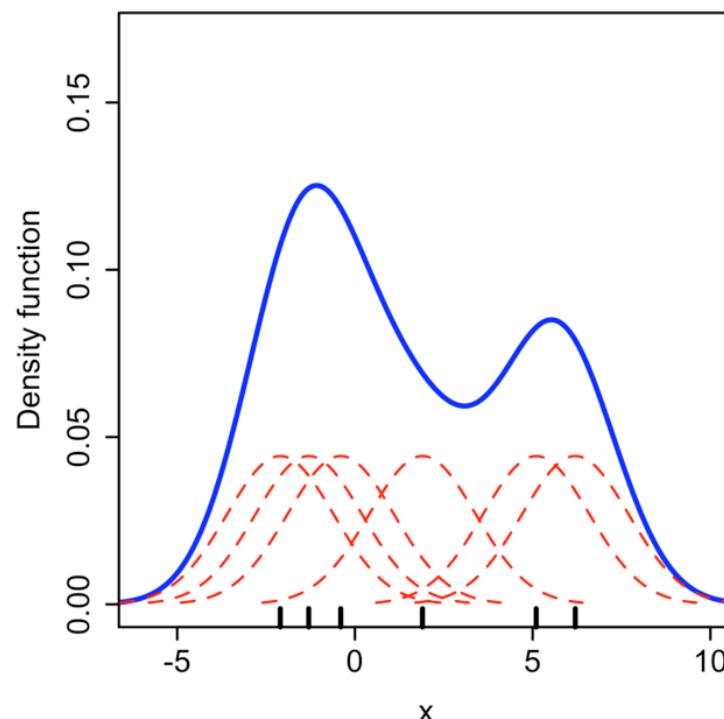
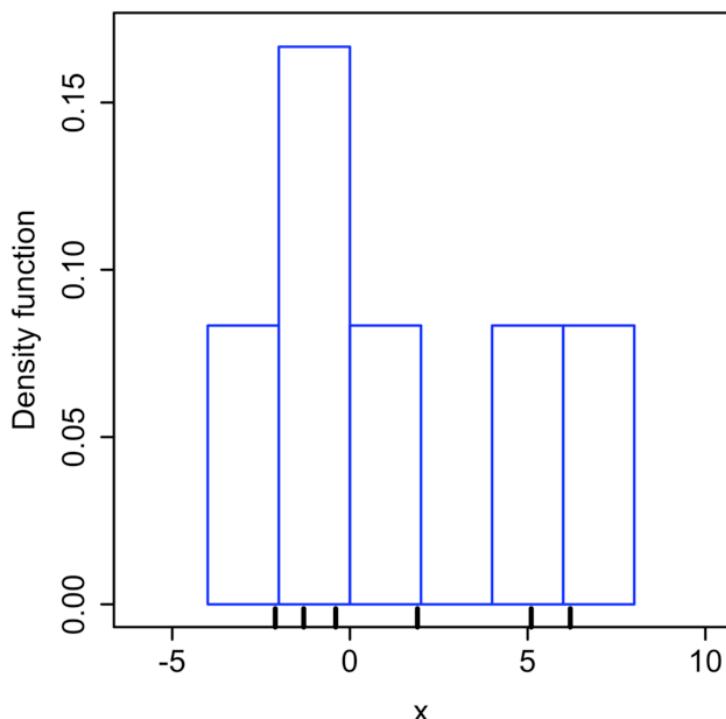
<https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d>



Kernel density estimation (KDE)



- A method for estimating the PDF of a random variable without assuming a parametric model
 - Some smoothness assumptions are required to guarantee convergence
- A smooth analog to histogram

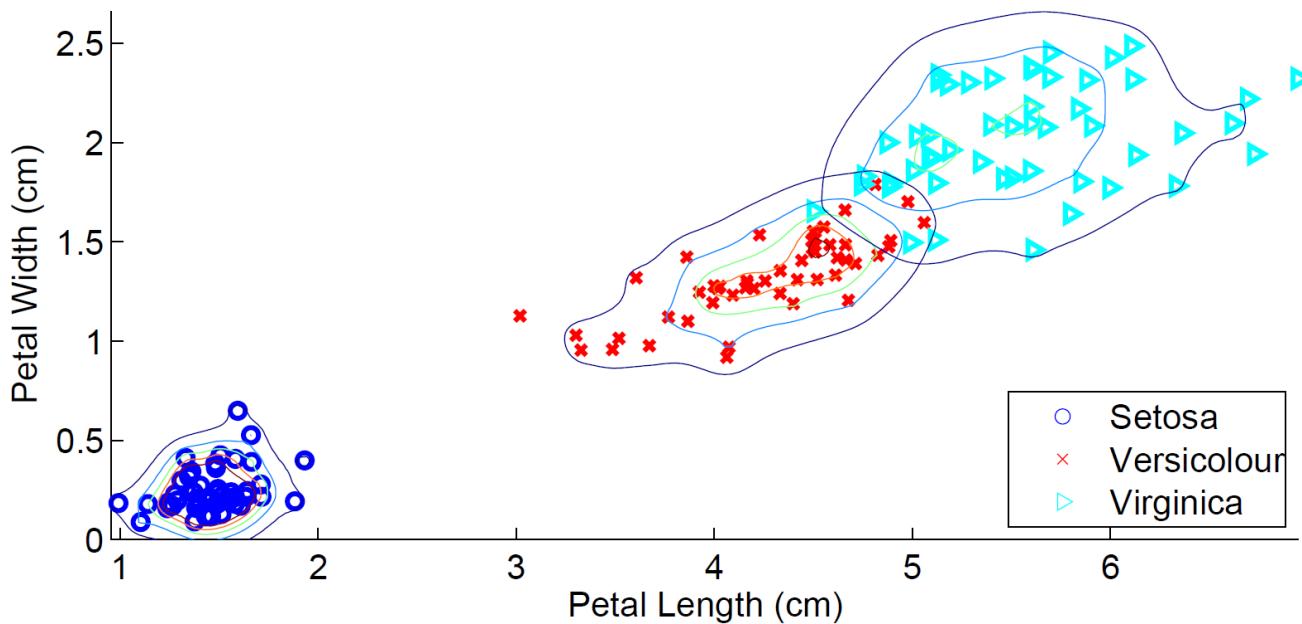


<https://commons.wikimedia.org/w/index.php?curid=57332968>



Information Theory in ML

Iris dataset (Fisher, 1936)



1. How similar are each of the classes to each other?
2. What is the intrinsic dimension of the data?
3. What is the best possible error rate any classifier can asymptotically achieve (i.e. the Bayes error)?
4. Are the variables dependent?
5. Are any of the points anomalies?
6. Which features are most relevant for classification?
7. How would you go about building a classifier?



Information Theory in ML



1. How similar are each of the classes to each other?
2. What is the intrinsic dimension of the data?
3. What is the best possible error rate any classifier can asymptotically achieve (i.e. the Bayes error)?
4. Are the variables dependent?
5. Are any of the points anomalies?
6. Which features are most relevant for classification?
7. How would you go about building a classifier?

Solution: Information Theory!



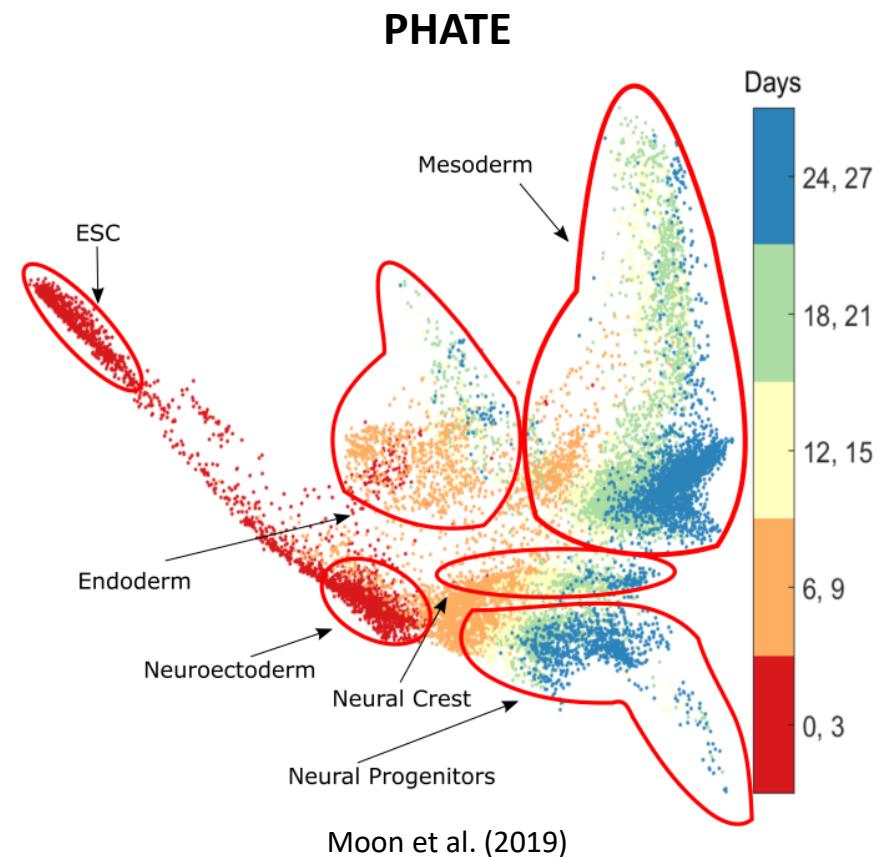
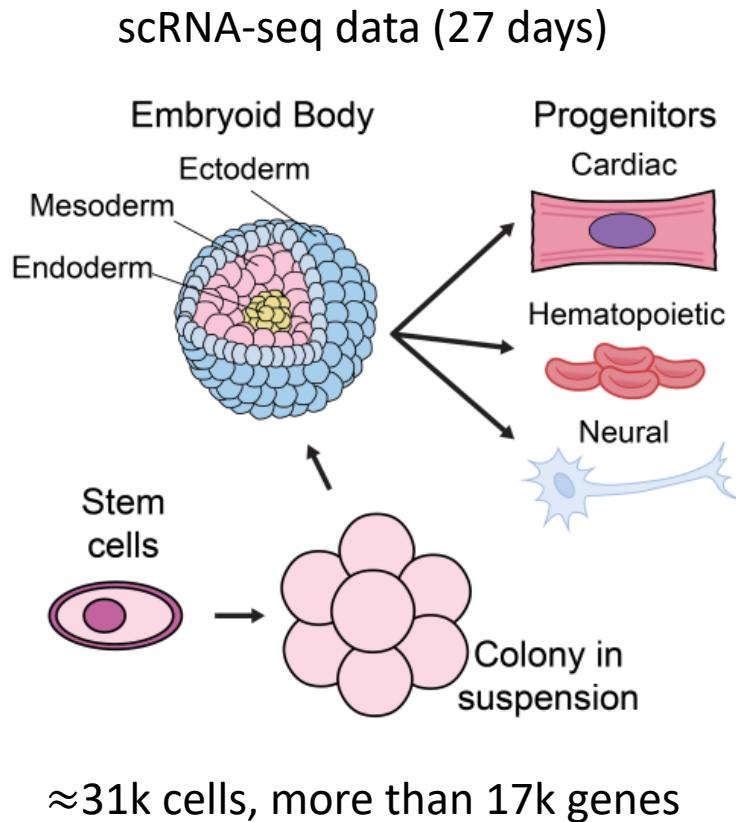
Dimensionality reduction

- High dimensional data suffers from the “curse of dimensionality”
 - Exponentially more samples are required for accurate results
- In practice, many high dimensional data are generated by a process that has only a few degrees of freedom
 - Ex: facial expressions (high dimensional images) are controlled by only a few muscles
- Goal is to reduce the dimension of the data
- **PCA** is the most common approach
 - Linear (a restrictive assumption) and relatively fast
- **Diffusion maps** is a nonlinear and robust dimensionality reduction method



PHATE

- A dimensionality reduction method designed for visualization
 - Preserves both local and global structure within the data

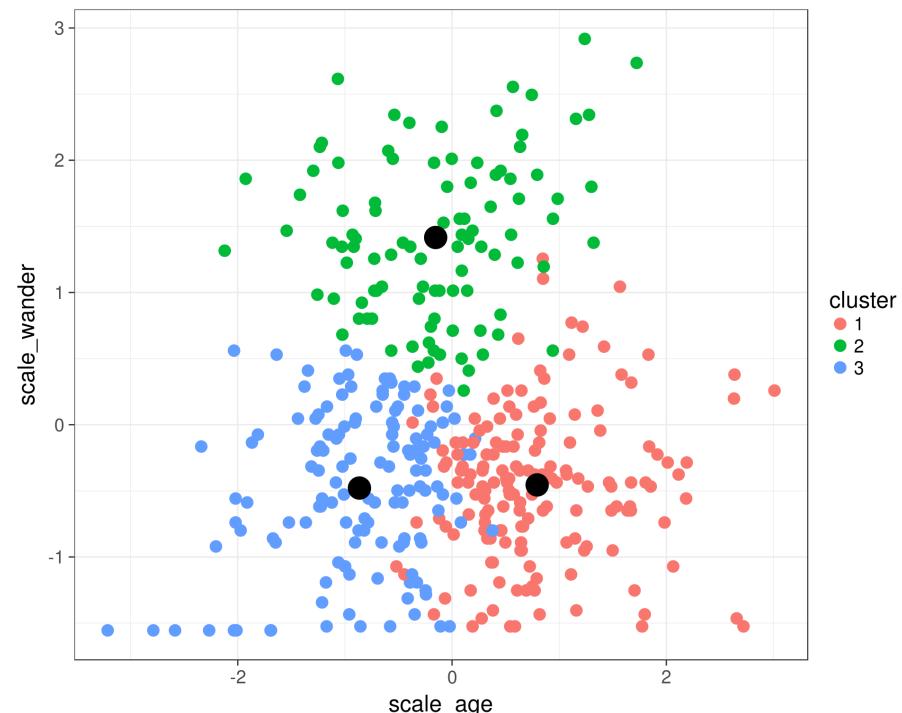




Clustering



- **Goal:** separate points into groups such that points within groups are “similar” to each other and points in different groups are “dissimilar”
 - Not well-defined \Rightarrow many clustering algorithms!
- A simple algorithm is **k-means**
 - Partitions the space into convex regions
 - Has many weaknesses
 - Many other methods essentially transform the input space and then apply k-means





Matrix completion/data imputation



- Not all entries in a data matrix are always observed
 - Ex: Netflix recommendation
 - Ex: Single-cell RNA-sequencing data
- Assumptions are imposed to learn the missing entries
 - Ex: people share viewing tastes
 - Ex: Many genes are co-expressed
- Linear (e.g. low-rank approximation) and nonlinear (e.g. MAGIC in van Dijk et al., 2018) methods exist

		-1		
			1	
1	1	-1	1	-1
1				-1
		-1		

1	1	-1	1	-1
1	1	-1	1	-1
1	1	-1	1	-1
1	1	-1	1	-1
1	1	-1	1	-1



Reinforcement learning



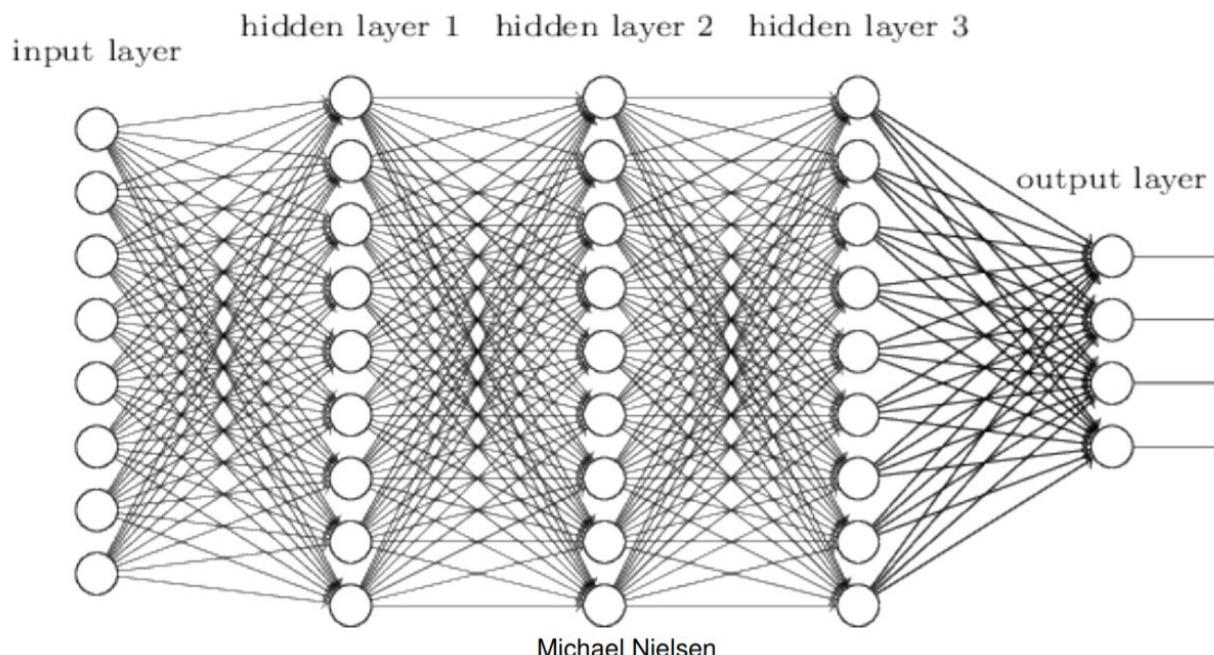
- **Goal:** given an agent within an environment and a set of possible actions (which may change depending on the current state within the environment), which action should the agent take to maximize its reward?
- **Ex:** Video games
 - Reward is to win (or not die or some measure of score)
- **Ex:** Controlling a robot
 - Reward would be not to crash or achieve some other objective





Neural Networks

- A network of nodes and edges with biases at each node and weights assigned to each edge
- Weights and biases are tuned via learning algorithms (i.e. optimization) to learn a function
 - Can be used for classification, regression, unsupervised learning, and reinforcement learning





Neural Networks



- Neural networks have become an important part of ML
 - Especially in image classification, language models, and generative models
- Deep learning refers to using a neural network with many layers
- We'll scratch the surface of neural networks in this class
- Take Deep Learning Theory and Applications (STAT 5685/6685) for more information
 - Taught by me

k-Nearest Neighbors Classifier



Binary Classification

- Training data

$$(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$$

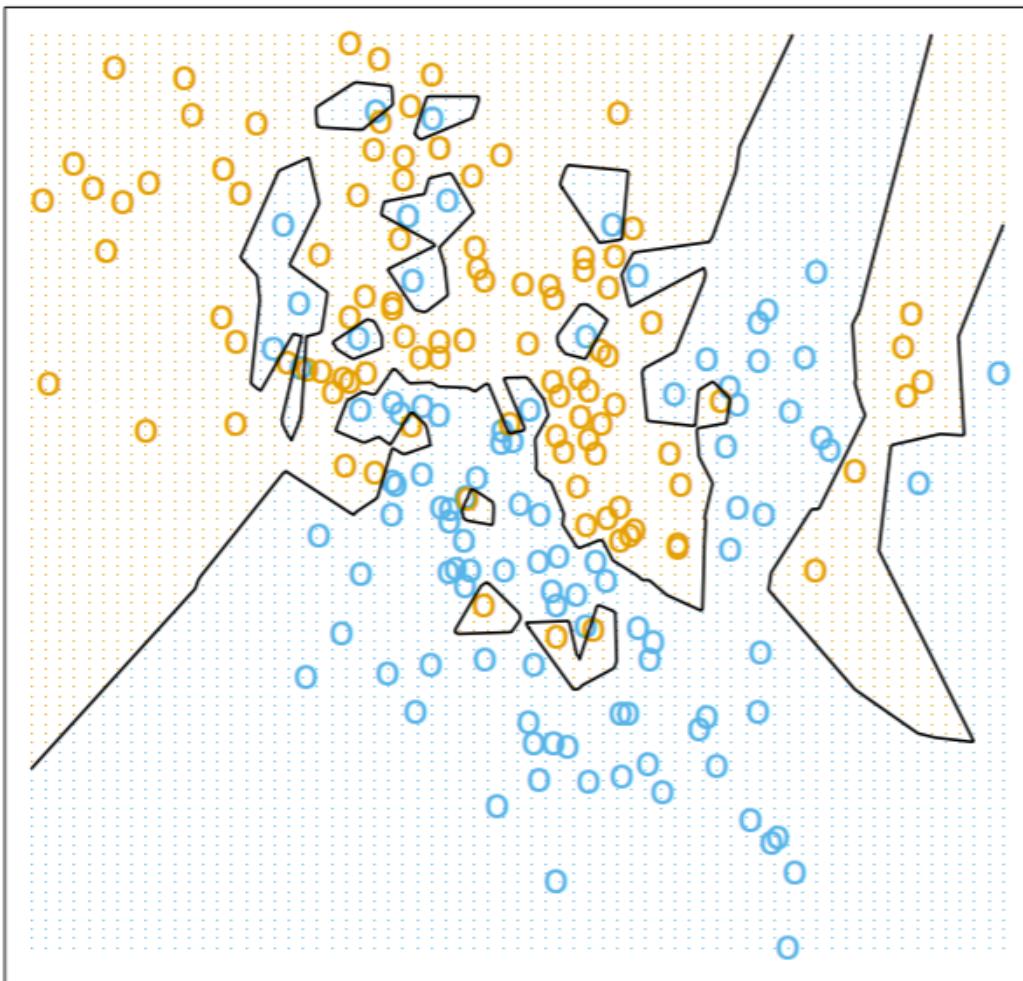
where $y_i \in \{-1, 1\}$.

- Test point \mathbf{x}
- A classifier is a function that maps \mathbf{x} to $\{-1, 1\}$
- A machine learning algorithm takes the training data as input and returns a classifier
- The nearest neighbor classifier is one of the simplest machine learning algorithms



Nearest Neighbor Classifier

- Assign x the same label as the closest training point x_i

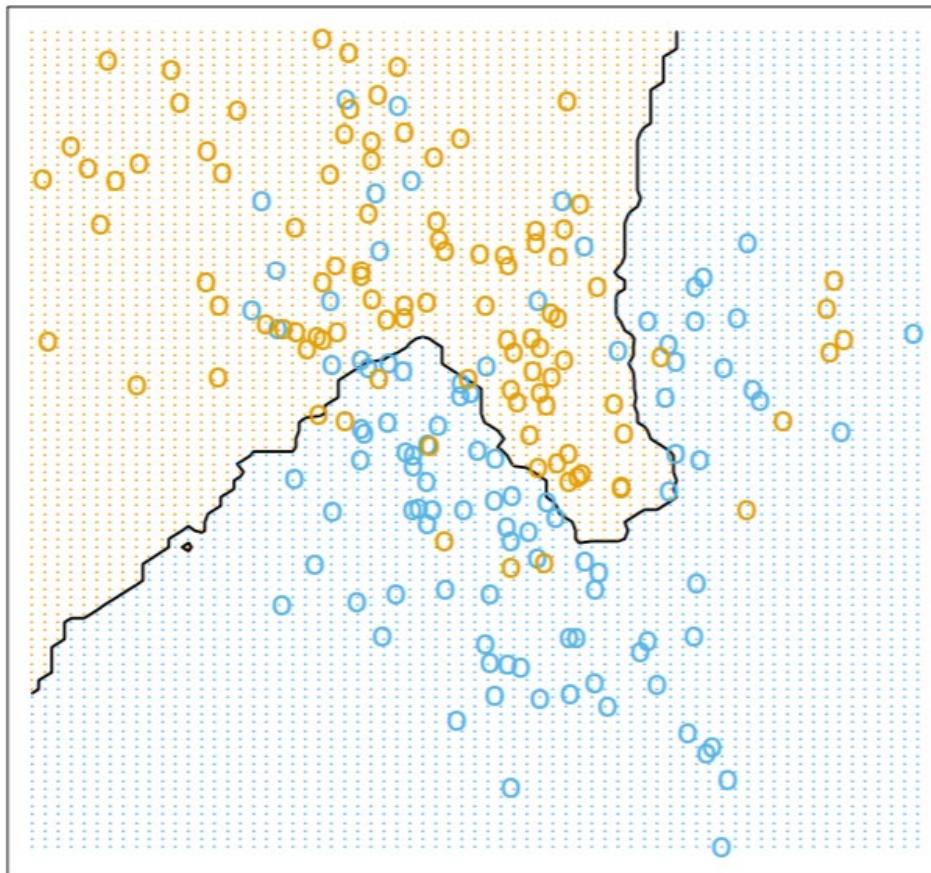




k-Nearest Neighbor Classifier

- Assign a label to x by taking the majority vote of the k closest training points x_i to x (k chosen to be odd)

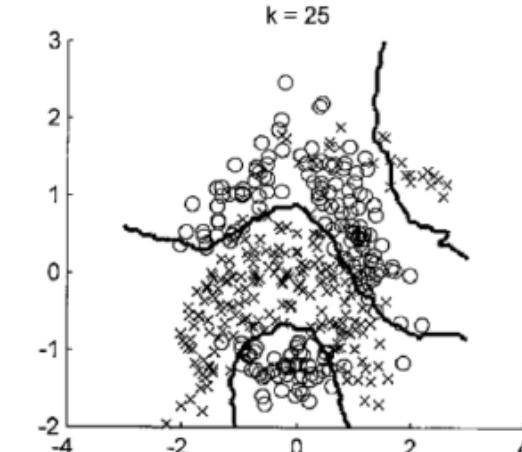
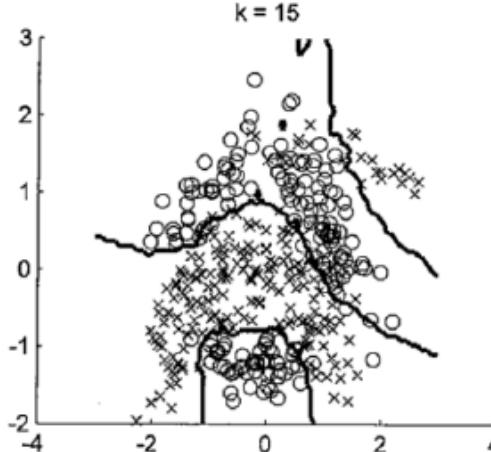
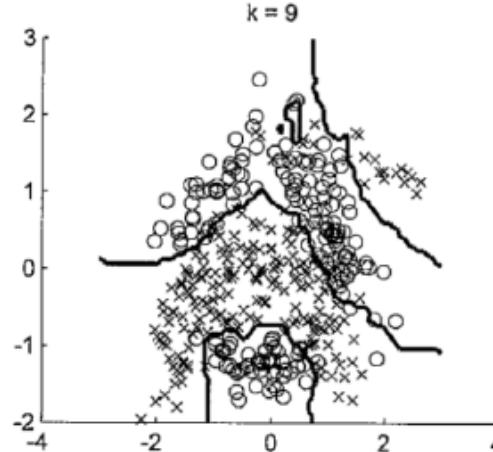
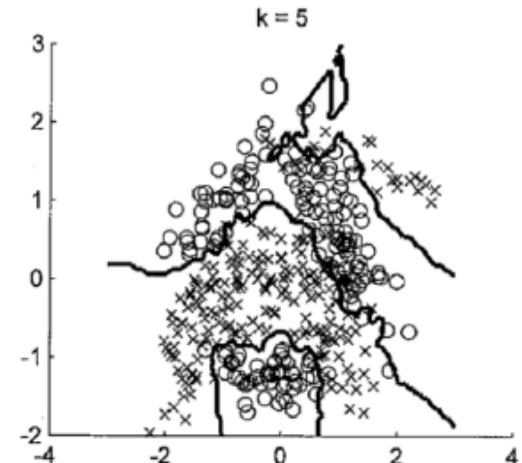
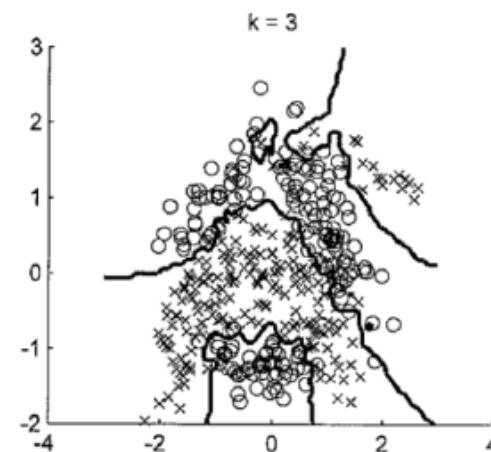
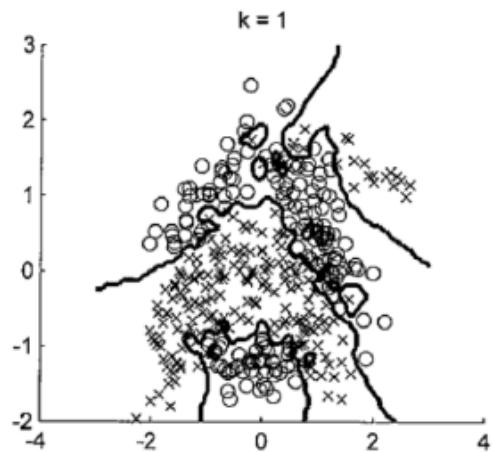
15-Nearest Neighbor Classifier





k-Nearest Neighbor Classifier

- Assign a label to x by taking the majority vote of the k closest training points x_i to x (k chosen to be odd)





Group Exercises

1. Is the k -NN classifier discriminative or generative?
Parametric or non-parametric? Linear or nonlinear?
2. What happens if k is chosen to minimize the training error (i.e. the error on the training data)?
3. What are some drawbacks of the k -NN classifier?
4. In what ways could the k -NN classifier be extended beyond the setting here?
5. How could k be chosen automatically?



Training vs. Test Error

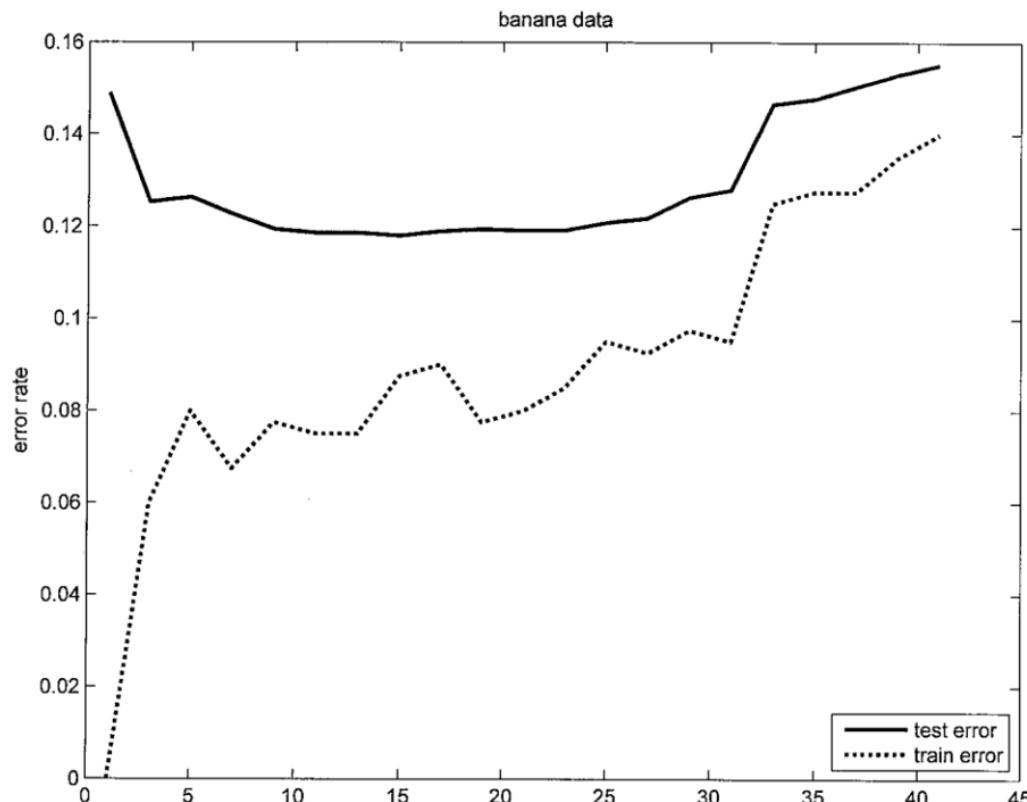
Some loose definitions (we'll tighten them up later):

- **Training error:** the fraction of incorrectly classified points in the training data
- **Test error:** the fraction of incorrectly classified points in a dataset that was not used for training (typically referred to as a **test set**)
- Typically, the total amount of available labeled data is split into two groups: a training set and a test set
 - Training error calculated from the training set
 - Test error calculated from the test set
 - We'll cover these “validation” approaches in more detail later
- Why do we need both errors?



Training Error is Optimistic

- k is a parameter that affects the smoothness/complexity of the classifier: larger k means more smoothness/less complexity
 - $\Rightarrow k$ controls a tradeoff between overfitting and underfitting





Further reading

- ISL Chapters 1 and 2
- ESL Chapter 1 and Section 2.3.2