Introduction to Machine Learning

COS 424/524, SML 302: Fundamentals of Machine Learning

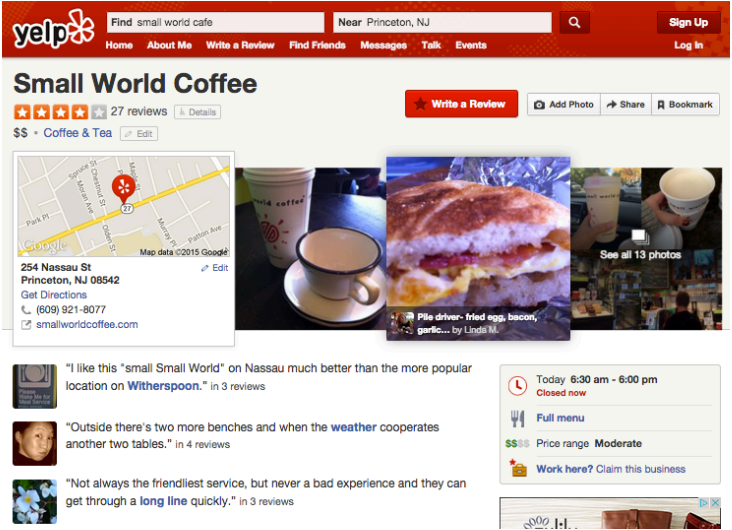
Professor Engelhardt

COS 424/524, SML 302

Lecture 1

Data are everywhere.

User ratings



Purchase histories

Delivered
Delivered on: Tuesday, December 16, 2014

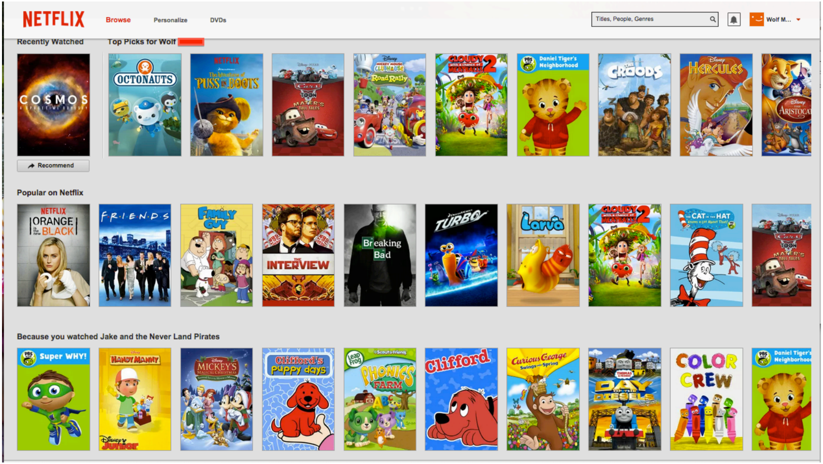
White Felt Cowgirl Cowboy Hat With Pink Sta
Sold by: Great Price Fast Service | Product que
$5.99
Condition: New - Brand New
Buy it Again

The Digi-Piggy White
Sold by: ToyBurg
$13.49
Condition: New - Brand new and factory seal
is #1. Over 7 years experience selling on Ama
Buy it Again

Alan Turing: The Enigma: The Book That Insp
Hodges, Andrew
Sold by: Amazon.com LLC
$15.01
Condition: New
Buy it Again

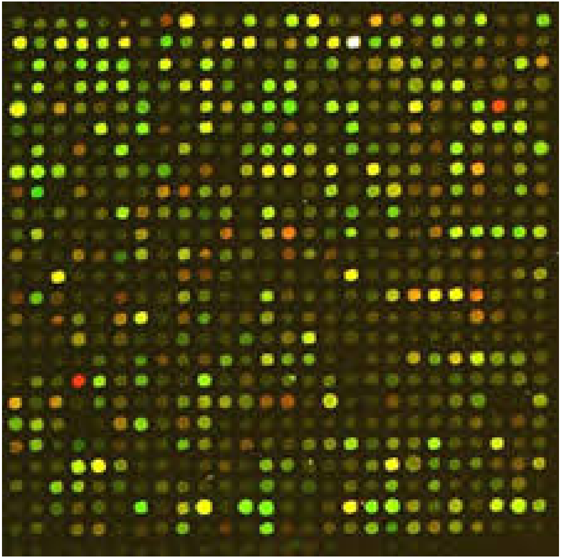
Melissa & Doug Princess Soft Toys 14" Plush
Sold by: Amazon.com LLC
$11.99
Condition: New
Buy it Again

Recommendations

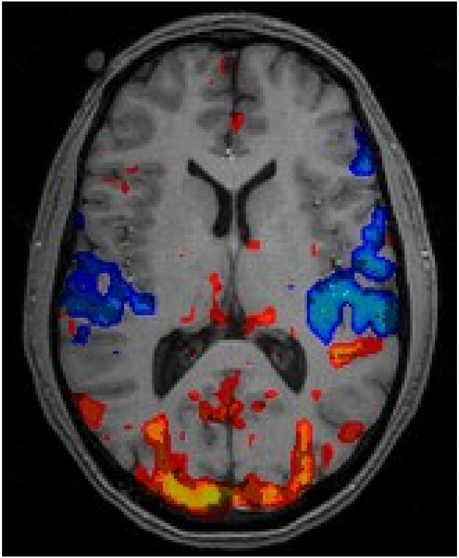


Document collections

Genomics



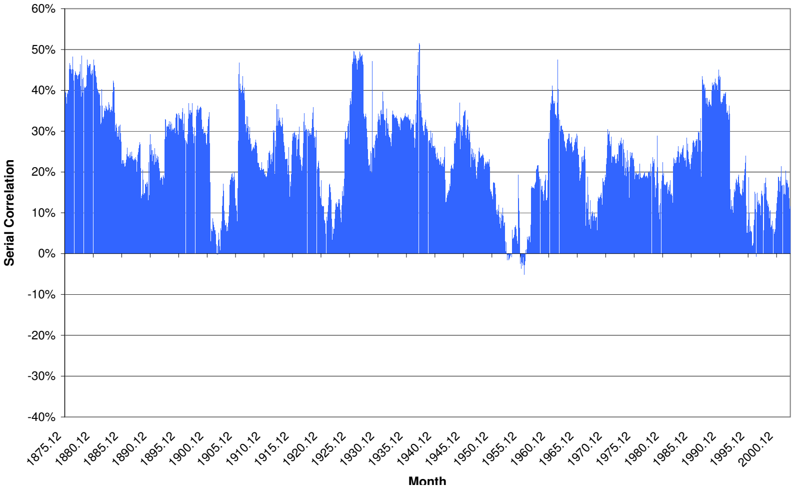
Neuroscience



Social networks

Astrophysics

Finance



First order autoregressive coefficients for monthly returns of the S&P Composite Index with five-year sliding window from Jan 1871 to Apr 2003.

Data can help us answer questions.

Will Netflix user #493234 like Transformers?

Group these images into 3 groups

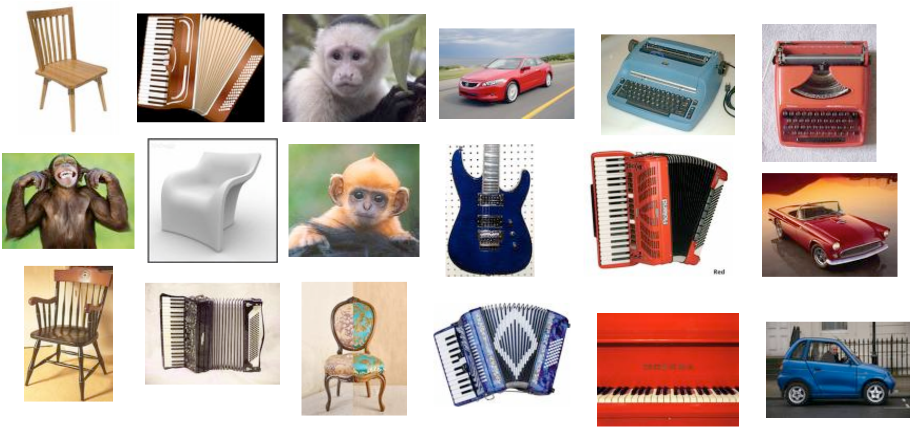








Rank these images...



- ...according to relevance to `instrument`.
- ...according to relevance to `machine`

Is this spam?

Subject: CHARITY.
Date: February 4, 2008 10:22:25 AM EST
To: undisclosed-recipients:;
Reply-To: s.polla@yahoo.fr

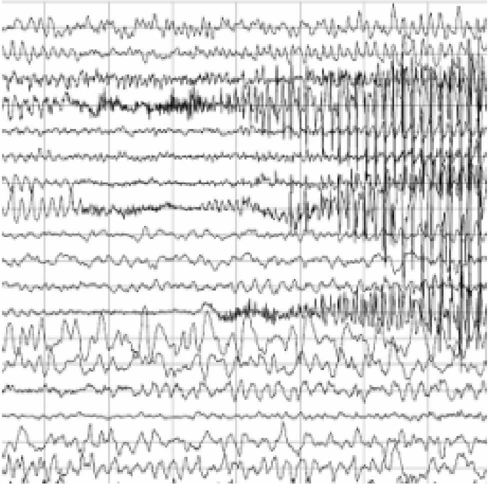
Dear Beloved,
My name is Mrs. Susan Polla, from ITALY. If you are a christian and interested in charity please reply me at : (s.polla@yahoo.fr) for insight.
Respectfully,
Mrs Susan Polla.

How about this one?

From: [snipped]
Subject: Superbowl?
Date: January 30, 2008 8:09:00 PM EST
To: blei@cs.princeton.edu, [snipped]

Anyone interested in coming by to watch the game? Beer and pizza, I'd imagine. If anyone wants, we could get together earlier, play a board game or cards or roll up characters or something. Takers?

When did the seizure begin?



Who was at graduation?

Did Neandertals and Homo sapiens share tools?

RESEARCH ARTICLE

A Draft Sequence of the Neandertal Genome

Richard E. Green,[1]*†‡ Johannes Krause,[1]†§ Adrian W. Briggs,[1]†§ Tomislav Maricic,[1]†§ Udo Stenzel,[1]†§ Martin Kircher,[1]†§ Nick Patterson,[2]†§ Heng Li,[2]† Weiwei Zhai,[3]†|| Markus Hsi-Yang Fritz,[4]† Nancy F. Hansen,[5]† Eric Y. Durand,[3]† Anna-Sapfo Malaspinas,[3]† Jeffrey D. Jensen,[6]† Tomas Marques-Bonet,[7,13]† Can Alkan,[7]† Kay Prüfer,[1]† Matthias Meyer,[1]† Hernán A. Burbano,[1]† Jeffrey M. Good,[1,8]† Rigo Schultz,[1] Ayinuer Aximu-Petri,[1] Anne Butthof,[1] Barbara Höber,[1] Barbara Höffner,[1] Madlen Siegemund,[1] Antje Weihmann,[1] Chad Nusbaum,[2] Eric S. Lander,[2] Carsten Russ,[2] Nathaniel Novod,[2] Jason Affourtit,[9] Michael Egholm,[9] Christine Verna,[21] Pavao Rudan,[10] Dejana Brajkovic,[11] Željko Kucan,[10] Ivan Gušic,[10] Vladimir B. Doronichev,[12] Liubov V. Golovanova,[12] Carles Lalueza-Fox,[13] Marco de la Rasilla,[14] Javier Fortea,[14]¶ Antonio Rosas,[15] Ralf W. Schmitz,[16,17] Philip L. F. Johnson,[18]† Evan E. Eichler,[7]† Daniel Falush,[19]† Ewan Birney,[4]† James C. Mullikin,[5]† Montgomery Slatkin,[3]† Rasmus Nielsen,[3]† Janet Kelso,[1]† Michael Lachmann,[1]† David Reich,[2,20]*† Svante Pääbo[1]*†

Neandertals, the closest evolutionary relatives of present-day humans, lived in large parts of Europe and western Asia before disappearing 30,000 years ago. We present a draft sequence of the Neandertal genome composed of more than 4 billion nucleotides from three individuals. Comparisons of the Neandertal genome to the genomes of five present-day humans from different parts of the world identify a number of genomic regions that may have been affected by positive selection in ancestral modern humans, including genes involved in metabolism and in cognitive and skeletal development. We show that Neandertals shared more genetic variants with present-day humans in Eurasia than with present-day humans in sub-Saharan Africa, suggesting that gene flow from Neandertals into the ancestors of non-Africans occurred before the divergence of Eurasian groups from each other.

changed parts of their genome with the ancestors of these groups.

Several features of DNA extracted from Late Pleistocene remains make its study challenging. The DNA is invariably degraded to a small average size of less than 200 base pairs (bp) (*21*, *22*), it is chemically modified (*21*, *23–26*), and extracts almost always contain only small amounts of endogenous DNA but large amounts of DNA from microbial organisms that colonized the specimens after death. Over the past 20 years, methods for ancient DNA retrieval have been developed (*21*, *22*), largely based on the polymerase chain reaction (PCR) (*27*). In the case of the nuclear genome of Neandertals, four short gene sequences have been determined by PCR: fragments of the *MC1R* gene involved in skin pigmentation (*28*), a segment of the *FOXP2* gene involved in speech and language (*29*), parts of the ABO blood group locus (*30*), and a taste receptor gene (*31*). However, although PCR of ancient DNA can be multiplexed (*32*), it does not allow the retrieval of a large proportion of the genome of an organism.

The development of high-throughput DNA sequencing technologies (*33*, *34*) allows large-scale, genome-wide sequencing of random pieces of DNA extracted from ancient specimens (*35–37*) and has recently made it feasible to sequence ge-

[1]Department of Evolutionary Genetics, Max-Planck-Institute for

Data contain patterns that can help us solve problems.

In this course, you will learn about methods that find and exploit patterns in data.

- These algorithms draw from statistics and machine learning.
- Applications of these algorithms are everywhere:
  - natural sciences
  - web technology
  - finance
  - social sciences
  - political sciences
  - engineering
  - medicine
  - politics
  - arts
  - commercial businesses
  - ...and all fields of study that have data.

In this course, you will learn about methods that find and exploit patterns in data.

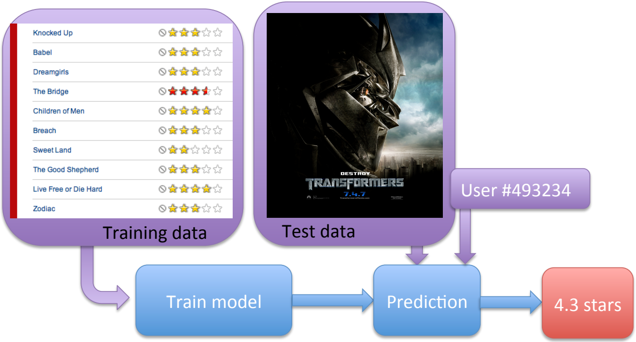
Your goal:

- Become fluent in thinking about modern data analysis problems.
- Be able to understand a new tool
- Be able to apply the tool correctly to data
- Be able to interpret the results appropriately

- We will learn about a suite of tools in modern data analysis:
  - when to use them
  - the assumptions they make about data
  - their capabilities and their limitations.
- We will learn a language & process for solving data analysis problems.

In this course, you will learn about methods that find and exploit patterns in data.

1. Collect, find, procure, or be given data.
2. Clean these data.
3. Analyze and explore the data to find patterns.
4. Use those patterns to accomplish a data analytic goal.



Patterns... in numbers

1. 1, 3, 5, 7, 9,...
2. 1, 2, 4, 8, 16,...
3. 1, 1, 2, 3, 5,...
4. 2, 5, 10, 17, 26,...
5. 14, 18, 23, 28, 34,...

Patterns... in numbers

1. 1,3,5,7,9,11,13,15: $(2n - 1)$
2. 1,2,4,8,16, 32, 64, 128: (2^{n-1})
3. 1,1,2,3,5,8,13,21: $(x_n = x_{n-2} + x_{n-1})$
4. 2,5,10,17,26,37,50,65: $(n^2 + 1)$
5. 14,18,23,28,34,42,50,59: (Stops on 1 NYC subway)

Patterns... in text

Fall, leaves, fall
Emily Bronte

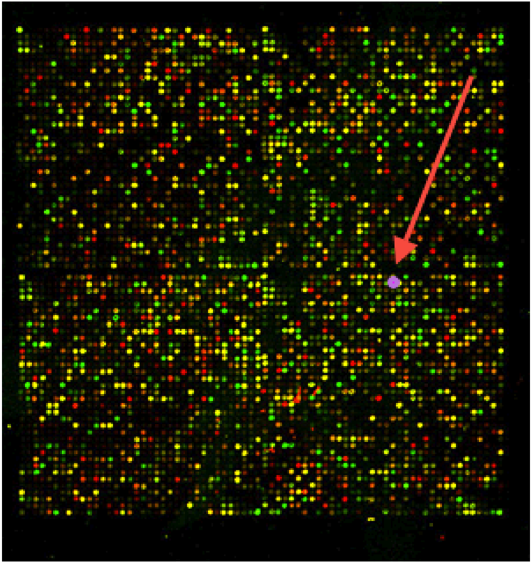
Fall, leaves, fall; die, flowers, away;
Lengthen night and shorten —;
Every leaf speaks bliss to me
Fluttering from the autumn —.
I shall smile when rings of snow
Blossom where the rose should —;
I shall sing when night's decay
Ushers in a drearier —.

Patterns... in text

Winter's a good time to stay in and cuddle,
But put me in summer and I'll be a ...

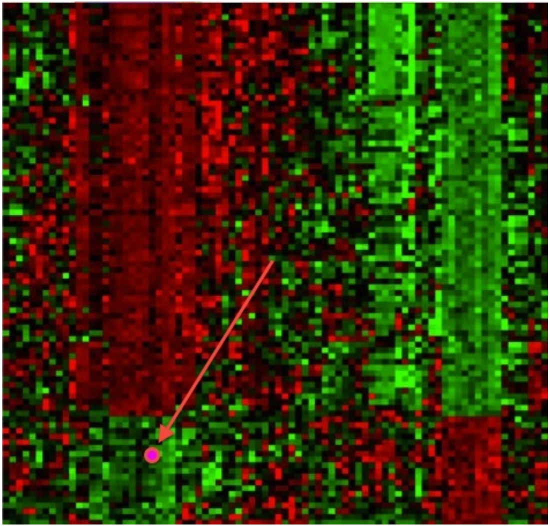
Patterns... in images

What is the value of the circled spot?



Patterns... in images

What is the value of the circled spot?



What do patterns help us do?

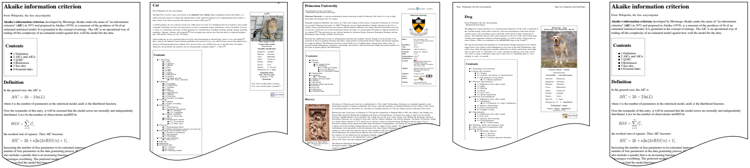
1. Predict
2. Classify
3. Cluster
4. Explore
5. Compress
6. Identify anomalies
7. Uncover associations
8. Discover latent structure

These are all tasks that we want to perform in data analyses.
Patterns make them possible.

How these ideas are organized

- Machine learning is broad; there are many ways to organize it.
- Fundamental idea: understand assumptions behind methods.
- We will use **statistical models** as an organizing principle.
  - Gives us a language for describing assumptions
  - Gives us a way of deriving and interpreting algorithms
- Some other recurring themes and concepts:
  - Supervised and unsupervised learning
  - Discrete data and continuous data
  - Ways of representing data
  - Adding sophistication to simple models
  - Building algorithms for efficiency

Understanding assumptions



- Methods make **assumptions** about the data.

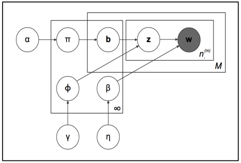
Example: modeling documents

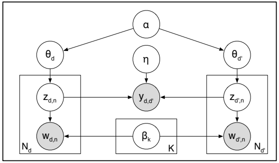
Documents are collections of words. They can be analyzed

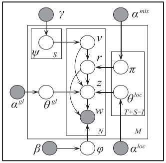
- as a sequence of words or as a "bag of words"
- independent of each other or correlated to each other

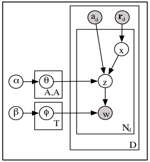
- What are the assumptions behind the methods?
- When and why are they useful? (Function of data and goals.)

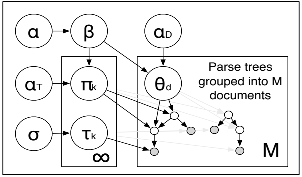
Graphical models

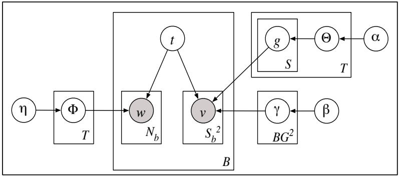


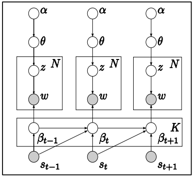






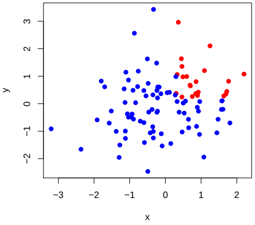






Graphical models represent assumptions in a statistical model.

Supervised methods



- **Supervised methods** find patterns in **fully observed** data and then try to predict unobserved variables in **partially observed** data.

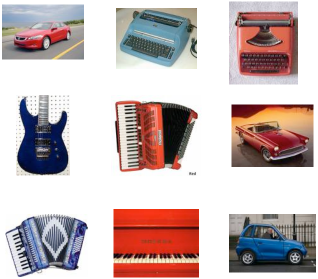
Example: email classification

We might observe a collection of emails that are categorized into *spam* and *not spam*.

After learning about what features indicate each category, we will automatically categorize new emails.

Unsupervised methods

Unsupervised methods find **hidden structure** in data that does not have a predefined label.



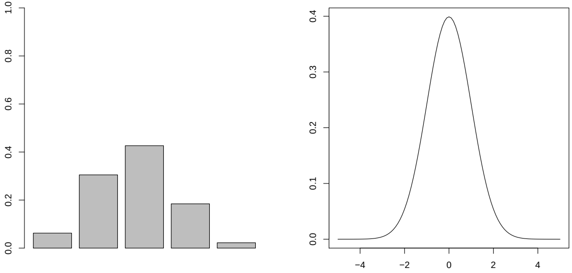
- Unsupervised learning is more difficult to evaluate than supervised learning.

Example: Image analysis

A museum has images of their collection that they want grouped by similarity of topic into fifteen different groups.

- These kinds of methods are widely used to find patterns that are not obvious in data.

Discrete and continuous methods



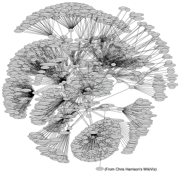
- *Discrete methods* manipulate a finite set of values
  - e.g., classification into one of five categories.
- *Continuous methods* manipulate continuous values
  - e.g., predict the change in stock value.

Data representation: Features

$\longrightarrow \langle 1.5, 3.2, -5.1, \ldots, 4.2 \rangle$

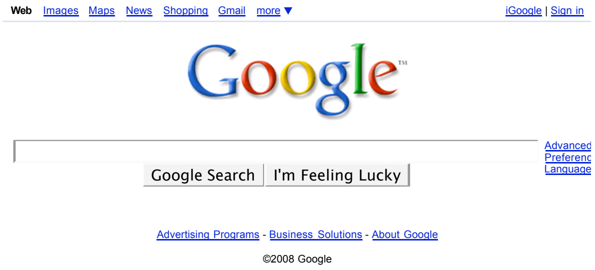
Republican nominee George Bush said he felt nervous as he voted today in his adopted home state of Texas, where he ended...

$\longrightarrow \langle 1, 0, 0, 0, 5, 0, 9, 3, 1, \ldots, 0 \rangle$



$$\longrightarrow \begin{bmatrix} 1 & 0 & 1 & \ldots & 0 \\ 0 & 1 & 1 & \ldots & 0 \\ 1 & 0 & 0 & \ldots & 1 \\ & \ldots & & & \\ 0 & 0 & 0 & \ldots & 0 \end{bmatrix}$$

Computational efficiency



- What we do with data is constrained by computation.
- We need to understand constraints and tailor our methods to them.
- We need to use methods that scale to the size of our data.

A statistician's manifesto (adapted)

(From T. Hastie, via J. McAuliffe and D. Blei)

- Understand the ideas behind the statistical methods, so you know how to use them, when to use them, and when *not* to use them.
- Complicated methods build on simple methods.
  Understand—and apply—simple methods first.
- The results of a method are of little use
  without an assessment of how well or poorly it is doing.

Ethics in Machine Learning

The field of ML is plagued with biased data, methods that amply these biases, and a homogeneous research community.

- As we discuss methods and assumptions, we will also discuss biases and questions of ethics.
- Homework assignments now focus exclusively on data with implications in social justice.
- I have made a targeted effort to include more diverse voices in the curriculum.

Logistics: course syllabus

Lecture	Week	Subject	Reading
01	01 Feb	Introduction	MLAPP Ch 1
02	01 Feb	Probability and statistics review	MLAPP Ch 2; [Opt] MLAPP Ch 3.1-3.4
03	08 Feb	Graphical models	MLAPP Ch 10.1-10.2, 10.4
04	08 Feb	Probabilistic classification	MLAPP Ch 3.5
05	15 Feb	Features and kernels	MLAPP 14.1-14.2
06	15 Feb	Kernel classifiers	MLAPP 14.3-14.5
07	22 Feb	Linear regression	MLAPP 7.1-7.3; [Opt] ESL Ch 3.1-3.2
08	22 Feb	Regularized linear regression	MLAPP 7.5.1,7.6.1,7.6.2; [Opt] ESL Ch 3.4
09	01 Mar	Logistic regression	MLAPP 8.1-8.2
10	01 Mar	Generalized linear models	MLAPP 9.1-9.3.2; [Opt] McCullagh and Nelder, Ch 2
11	08 Mar	K-Means	MLAPP 11.1-11.3
12	08 Mar	Mixture models	
	15 Mar	Spring break	
13	17 Mar	Optimization	MLAPP 8.3 & 8.5
14	22 Mar	Expectation-maximization	MLAPP 11.4-11.6
15	22 Mar	Hidden Markov models	MLAPP 17.1-17.2
16	29 Mar	Dimension reduction and PCA	MLAPP Ch 12.1-12.2
17	29 Mar	Factor analysis	
18	05 Apr	Probabilistic topic models	Blei (2011)
19	05 Apr	Communities in networks	Airoldi et al. (2008)
20	12 Apr	Dirichlet processes	MLAPP 25.2
21	12 Apr	Gaussian process regression	Roberts et al. 2013
22	19 Apr	Markov chain Monte Carlo	MLAPP 23 (optional), 24.1-24.3.5
23	19 Apr	Scalable machine learning	MLAPP 21.1-21.5 (not 21.4)
24	26 Apr	Summary and discussion	
	M 03 May	Poster session	
	W 05 May	Dean's Date	Projects due (5pm EST)

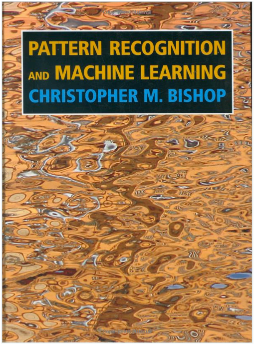
Course requirements

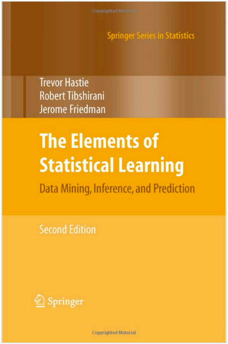
- Read the material and take 10 of 12 weekly online reading quizzes (10% of your grade)
- Do the homework (60% of your grade).
- Prepare a final project (30% of your grade).

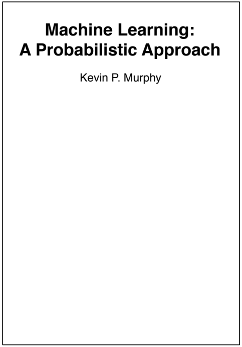
Highly-recommended prior knowledge

- Linear algebra (e.g., Math 202)
- Intro statistics or probability (e.g., SML 201)
- Programming (e.g., COS 126)

Course reading







- We will provide reading materials, mostly from these books.
- There will be a weekly reading quiz (ten of 12).
- Online quiz must be completed before 9pm EST Thursday.
- The link to the online quiz is in Canvas.

Homeworks: mini-projects

- Three different three week exercises
- To be done alone or in pairs.
- To hand in: five page write up on your analysis of a data set
    - Homework 1: Classify tweets into supportive or against #BLM and police
    - Homework 2: Predict characteristics of 5,000 low income teenagers using survey data
    - Homework 3: Identify patterns in police complaint networks
- We will post an 'example homework' on Canvas;
- Use our LaTeX template for your homework write ups.
- First homework is available on Canvas now.

Final Project

- The final project is for you to stretch your data analytic wings
- ... on an applied data analysis project
- You may work alone or in groups of up to four people.
- You may:
  - Extend analysis of a homework data set with new & exciting method
  - Analyze new data with methods you discovered or developed
- Project teams and proposals are due two days after the final assignment is due;
- The course will culminate in a virtual poster session the Monday before Dean's Date;
- The eight page write up of your project due by 5pm on Dean's Date.

Course staff

- **Professor**
  Barbara Engelhardt
  OH: Thursdays, 1:00PM-2:30PM
  bee@princeton.edu
- **Lecturer**
  Dr. Xiaoyan Li
  xiaoyan@princeton.edu
- **Teaching Assistants**
  To be announced
  OH: See Canvas

Communicating with Canvas

We will use Canvas to manage all communication

- Most questions answered within one business day
- Host discussions among yourselves
- Use for any kind of technical question
- Use for most administrative questions
- Can use to send us private questions too, in lieu of email
- We will distribute lecture slides and homework instructions on Canvas.

A brief note on the honor code

With COS courses comes the issue of honor code violations.
We define an honor code violation in this course as in the undergraduate handbook; this definition includes, but is not limited to

- using more than four words of text from a source without citing that source (including online sources);
- borrowing ideas without proper attribution;
- sharing work, including code, without proper attribution.

We have a simple approach to honor code violations in this course:

- we use software to determine if your code is similar to anyone else's code in the course or in any previous version of the course;
- we use software to determine if your homework and project papers borrowed text from any published or online sources, or each other;
- we address violations of the honor code quickly.

Items to do as soon as possible

- Sign up for Canvas website
- Fill out course survey (on Canvas)
- Download and start to work on Twitter analysis corpus

Additional Resources

All lecture slides will include a final ‘Resources’ slide, which includes many other sources of information on the ideas in lecture

Additional Resources: Get Involved

- Apply to the **Undergraduate and Graduate Certificate Program in Statistics and Machine Learning**:

  `sml.princeton.edu`

- Sign up for the email list on ML-stat talks at Princeton: `https://lists.cs.princeton.edu/mailman/listinfo/ml-stat-talks`
- Begin independent research!

Additional Resources: Read about ML

- *Machine Learning: A Probabilistic Approach* Murphy (2012).
- *An Introduction to Statistical Learning* James, Witten, Hastie, Tibshirani (2017) (available online).
- *The Elements of Statistical Learning* Hastie, Tibshirani, Friedman (2001) (available online).
- *Probabilistic Graphical Models: Principles and Techniques* Koller (2009).
- *Pattern Recognition and Machine Learning* Bishop (2006).
- *Machine Learning* Mitchell (1997).
- *Pattern Classification* Duda, Hart, Stork (2001).

- Journal Articles: Journal of Machine Learning Research (JMLR).
- Conference proceedings: NeurIPS, AI-STATS, ICML, UAI, ICLR.

Additional Resources

Listen and watch:

- This Week in Machine Learning (podcast)
- The Talking Machine (podcast)
- The Radical AI Podcast (podcast)

Start exploring interesting data:

- Data are everywhere! Search and ask.
- Machine Learning for Social Justice, collection of data sets: `https://sites.google.com/view/ml4sj/data`
- Kaggle.com has a large number of data sets with varying ease of use.