CSE 5523: Machine Learning and Statistical Pattern Recognition

Instructor: Alan Ritter

Administrative Details

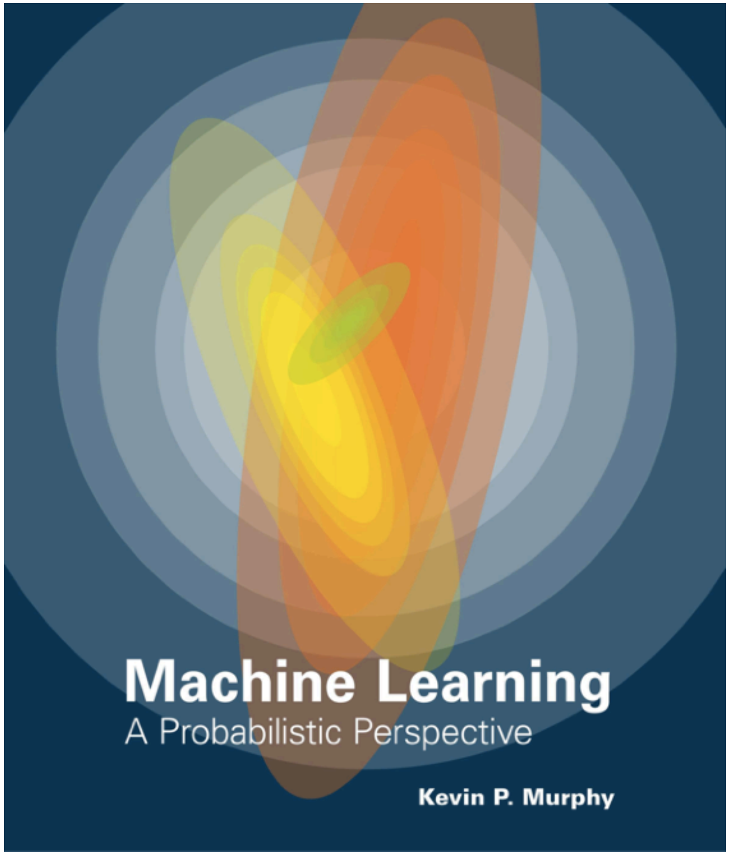
- Course Webpage
  - **http://aritter.github.io/courses/5523_fall18.html**
- Instructor
  - Alan Ritter
  - Office Hours: Tuesdays 4-5pm, Dreese 595
- TA
  - Shi Zong
  - Office Hours: Wednesdays 4-5pm Dreese 176

Administrative Details

- Piazza (discussion and resources)
  - https://piazza.com/class/ix8mu2qajn75mb
- Carmen (homework submission)
  - https://osu.instructure.com/courses/14167

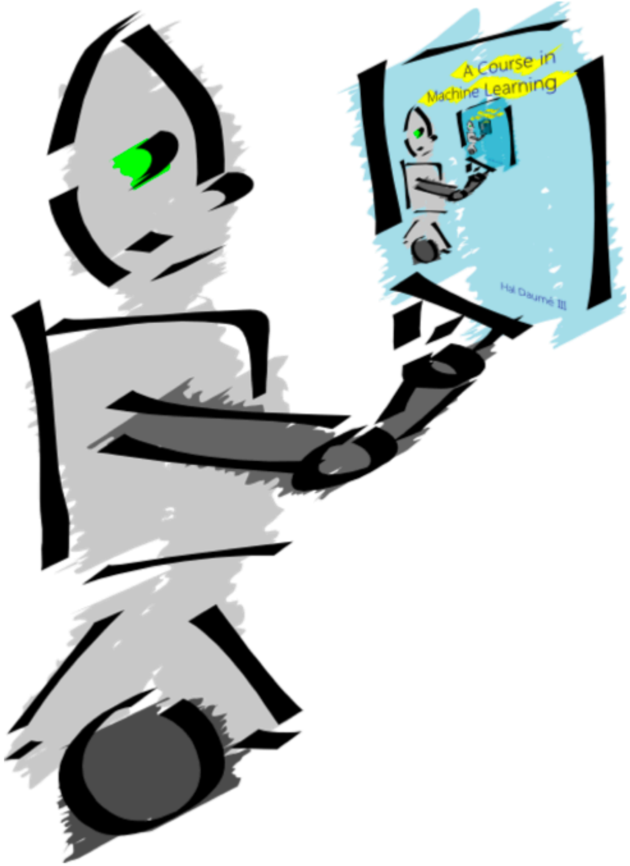
Reading

- Book
  - Kevin Murphy's ML book
    - https://www.cs.ubc.ca/~murphyk/MLbook/



Reading

- H. Daume III, A Course in Machine Learning
  - Free, online!
  - http://ciml.info/
  - (but incomplete…)



Prerequisites

- This class assumes you know:
  - Probability
  - Linear Algebra
  - Multivariate Calculus
  - Python (or have ability to learn Python quickly)
  - Numpy/scipy
  - Linux/Unix scripting
    - For windows users: https://www.cygwin.com/)
    - Windows Subsystem for Linux
    - VirtualBox
    - etc…

Prerequisites

- This class assumes you know:
  - Probability
  - Linear Algebra
  - Multivariate Calculus
  - Python (or have ability to learn Python quickly)
  - Numpy/scipy
  - Linux/Unix scripting
    - For windows users: https://www.cygwin.com/
    - Windows Subsystem for Linux
    - VirtualBox
    - etc…
- **Homework #1 is out**
  - Due next week
  - Turn in at beginning of class

Evaluation

Grading

Grading will be based on:

Participation (10%)

You will receive credit for asking and answering thoughtful questions related to the homework on Piazza and engaging in class discussion.

Homeworks (50%)

The homeworks will include both written and programming assignments. Homework should be submitted to the Dropbox folder in Carmen by 11:59pm on the day it is due (unless otherwise instructed). Each student will have 3 flexible days to turn in late homework throughout the semester. As an example, you could turn in the first homework 2 days late and the second homework 1 day late without any penalty. After that you will loose 20% for each day the homework is late. Please email your homework to the instructor in case there are any technical issues with submission.

Midterm (20%)

There will be an in-class midterm on Ocbober 30.

Final Projects (20%)

The final project is an open-ended assignment, with the goal of gaining experience applying the techniques presented in class to real-world datasets. Students should work in groups of 3-4. It is a good idea to discuss your planned project with the instructor to get feedback. The final project report should be 4 pages and is due on April 30. The report should describe the problem you are solving, what data is being used, the proposed technique you are applying, how you plan to evaluate your solution in addition to what baseline is used to compare against.

What to Expect

- Lots of math and programming
- Machine learning algorithms often difficult to debug
  - Need to think creatively about simple test cases.
  - We ***strongly*** recommend you start early.

- Questions?

A Few Quotes

- “A breakthrough in machine learning would be worth ten Microsofts” (Bill Gates, Chairman, Microsoft)

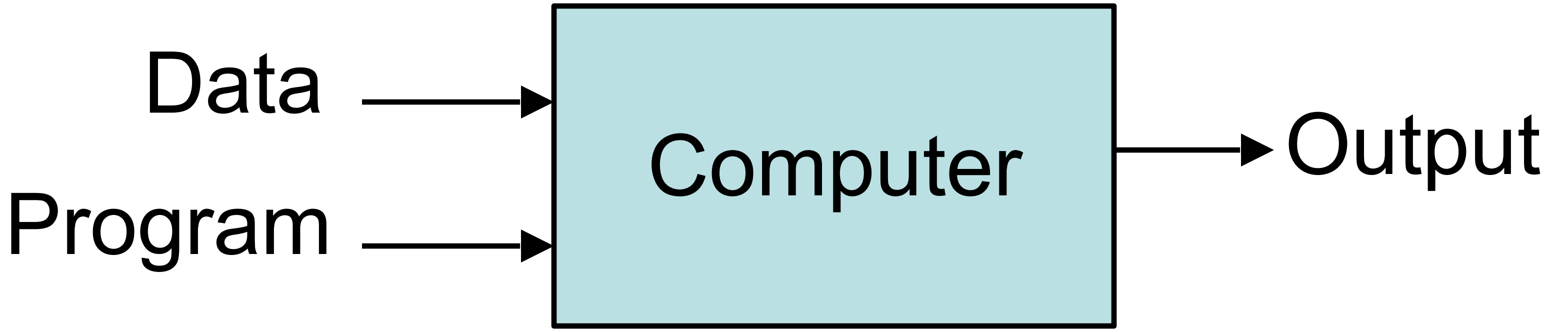
A Few Quotes

- “Machine learning is the next Internet” (Tony Tether, Director, DARPA)

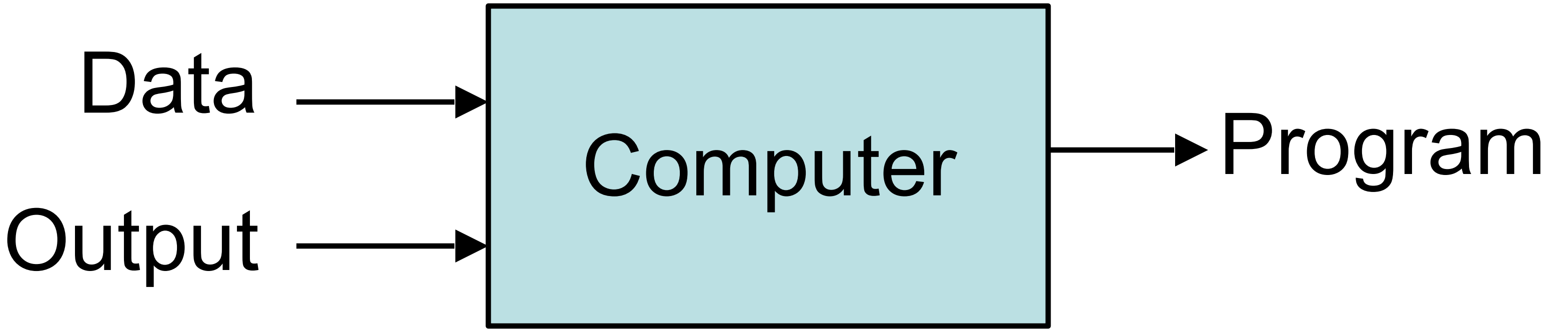
So What Is Machine Learning?

- Automating automation
- Getting computers to program themselves
- Writing software is the bottleneck
- Let the data do the work instead!

Traditional Programming



Machine Learning

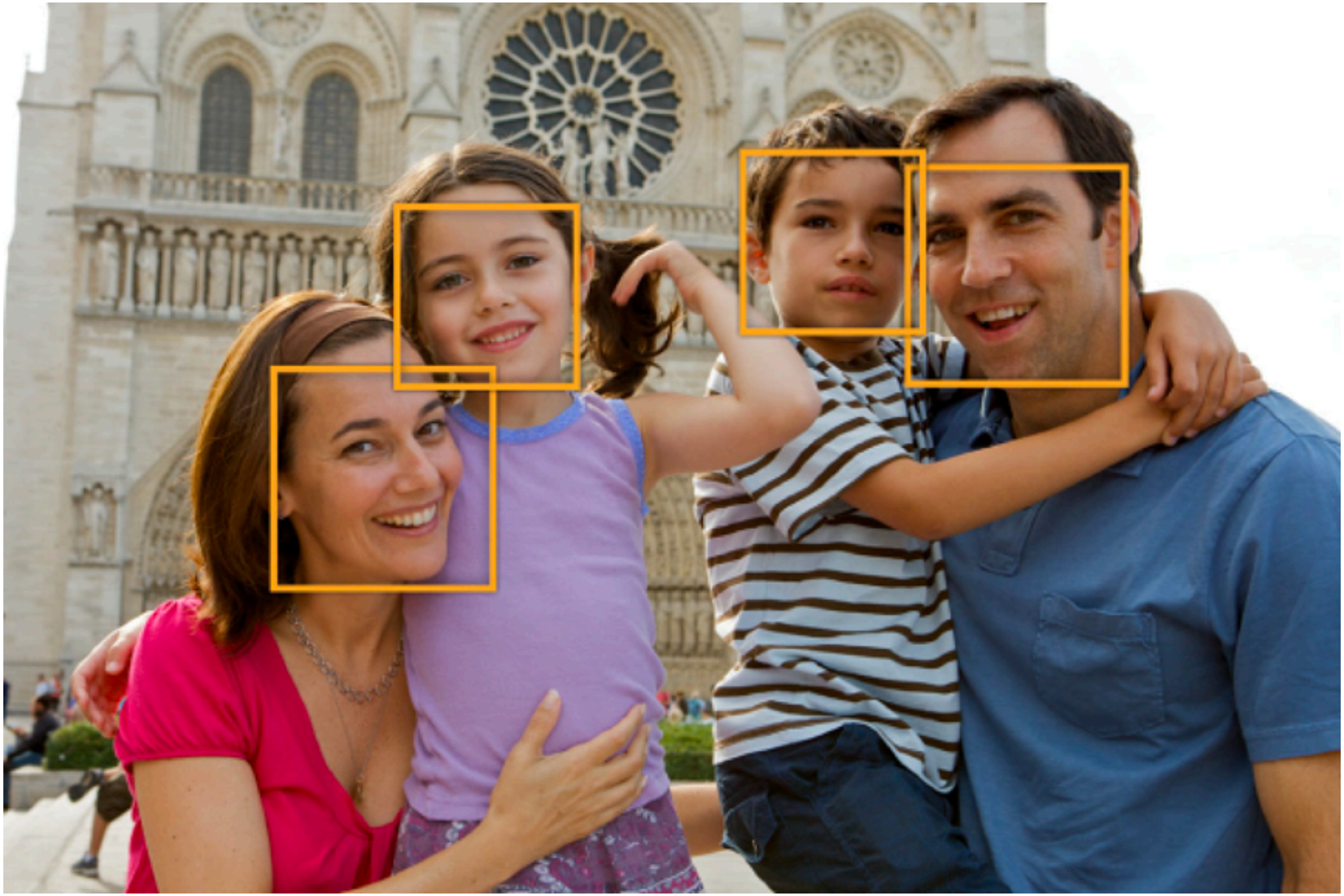


Magic?

No, more like farming

- Seeds = Learning Algorithms
- Nutrients = Data
- Farmer = You
- Plants = Programs

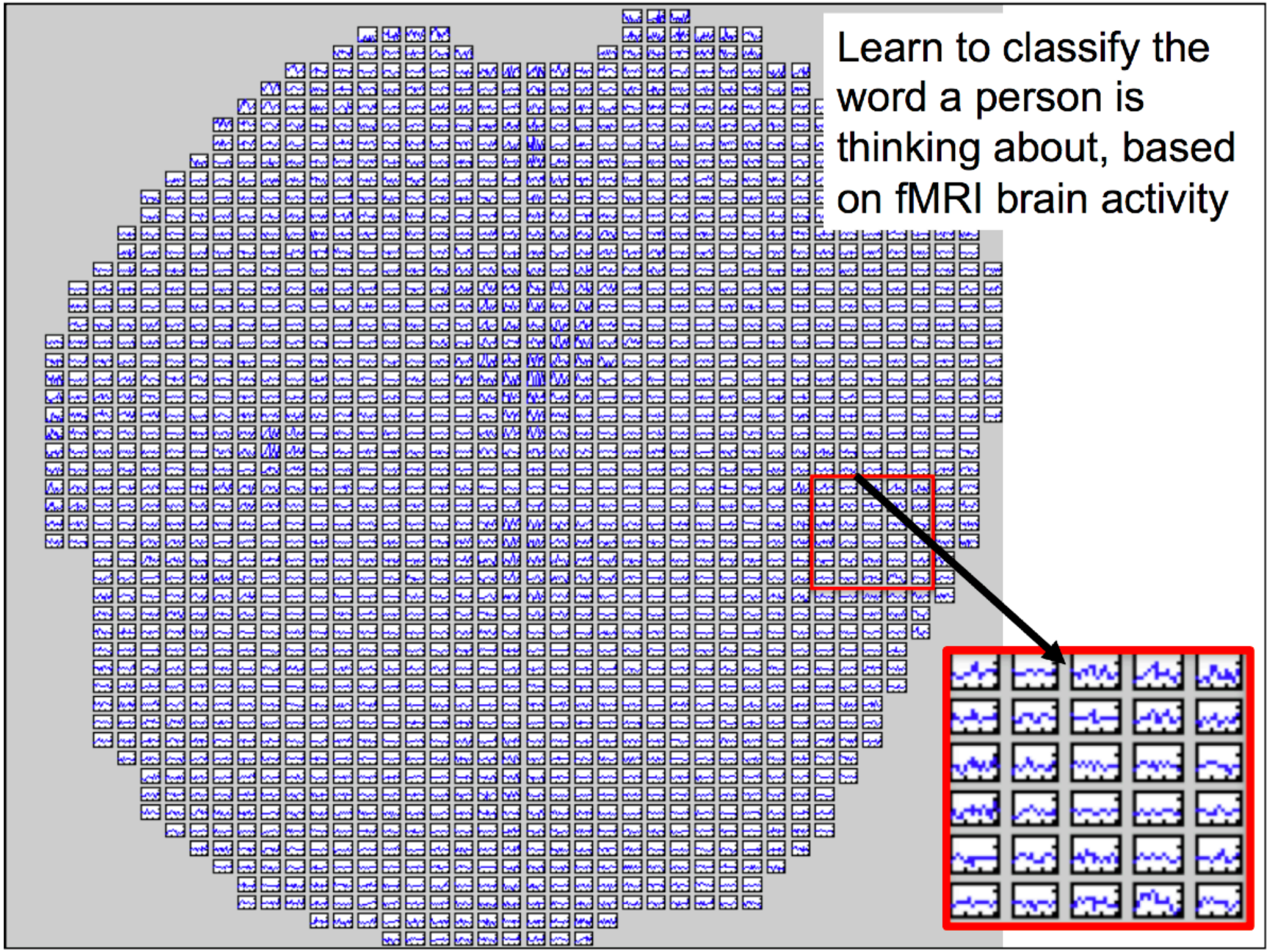
Sample Applications



Sample Applications

Sample Applications

Sample Applications



Machine Learning - Theory

PAC Learning Theory
(supervised concept learning)

examples (*m*)

representational complexity (*H*)

error rate (ε)

failure probability (δ)

$$m \geq \frac{1}{\epsilon}(\ln |H| + \ln(1/\delta))$$

Other theories for

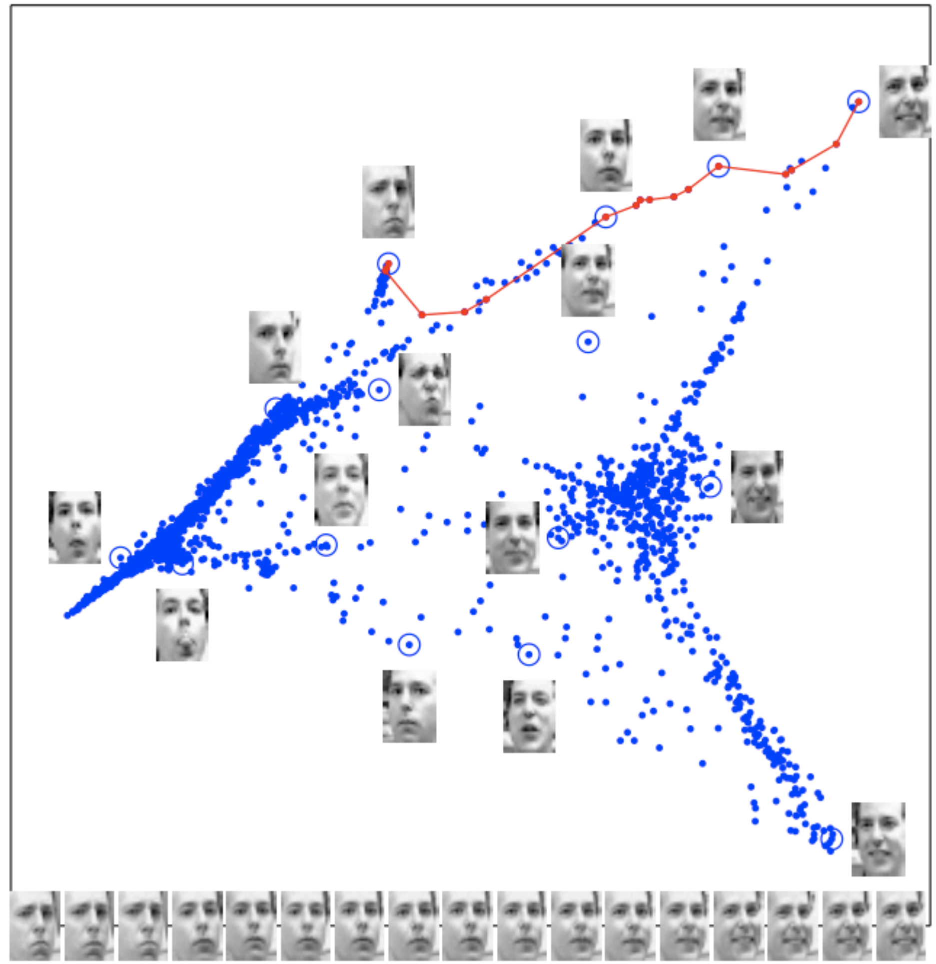
- Reinforcement skill learning
- Semi-supervised learning
- Active student querying
- …

… also relating:

- # of mistakes during learning
- learner's query strategy
- convergence rate
- asymptotic performance
- bias, variance

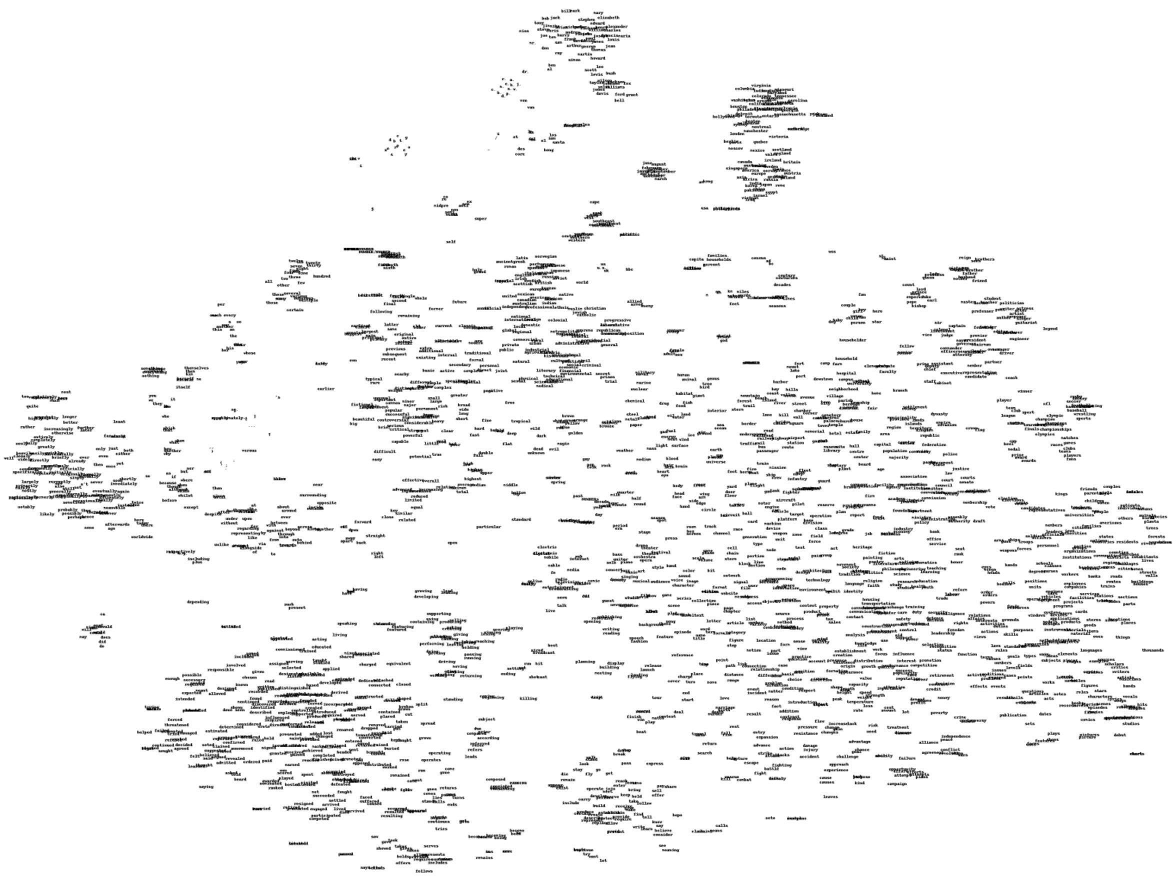
Embedding images

- Images have thousands or millions of pixels.
- Can we give each image a coordinate, such that similar images are near each other?



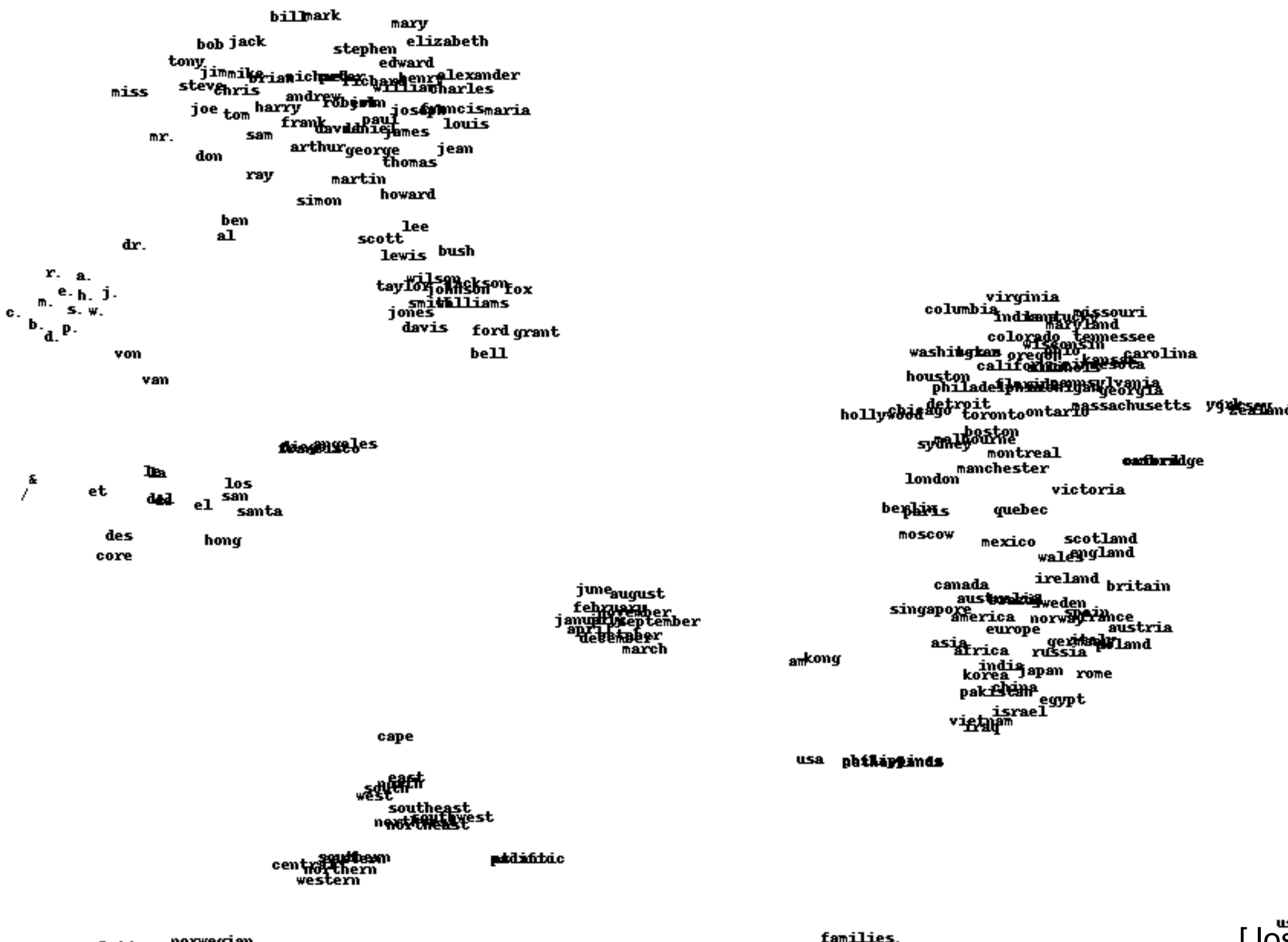
[Saul & Roweis '03]

Embedding words



[Joseph Turian]

Embedding words (zoom in)



[Joseph Turian]

Growth of Machine Learning

- Machine learning is preferred approach to
  - Speech recognition, Natural language processing
  - Computer vision
  - Medical outcomes analysis
  - Robot control
  - Computational biology
  - Sensor networks
  - …
- This trend is accelerating
  - Improved machine learning algorithms
  - Improved data capture, networking, faster computers
  - Software too complex to write by hand
  - New sensors / IO devices
  - Demand for self-customization to user, environment

Supervised Learning: find f

- Given: Training set $\{(x_i, y_i) \mid i = 1 \ldots n\}$
- Find: A good approximation to $f : X \rightarrow Y$

Examples: what are X and Y ?

- Spam Detection
  - Map email to {Spam,Ham}
- Digit recognition
  - Map pixels to {0,1,2,3,4,5,6,7,8,9}
- Stock Prediction
  - Map new, historic prices, etc. to $\Re$ (the real numbers)

Example: Spam Filter

- Input: email
- Output: spam/ham
- Setup:
  – Get a large collection of example emails, each labeled “spam” or “ham”
  – Note: someone has to hand label all this data!
  – Want to learn to predict labels of new, future emails

- Features: The attributes used to make the ham / spam decision
  – Words: FREE!
  – Text Patterns: $dd, CAPS
  – Non-text: SenderInContacts
  – …

Dear Sir.

First, I must solicit your confidence in this transaction, this is by virture of its nature as being utterly confidencial and top secret. …

TO BE REMOVED FROM FUTURE MAILINGS, SIMPLY REPLY TO THIS MESSAGE AND PUT "REMOVE" IN THE SUBJECT.

99 MILLION EMAIL ADDRESSES
FOR ONLY $99

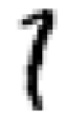
Ok, Iknow this is blatantly OT but I'm beginning to go insane. Had an old Dell Dimension XPS sitting in the corner and decided to put it to use, I know it was working pre being stuck in the corner, but when I plugged it in, hit the power nothing happened.

Example: Digit Recognition

- Input: images / pixel grids
- Output: a digit 0-9
- Setup:
  - Get a large collection of example images, each labeled with a digit
  - Note: someone has to hand label all this data!
  - Want to learn to predict labels of new, future digit images

- Features: The attributes used to make the digit decision
  - Pixels: (6,8)=ON
  - Shape Patterns: NumComponents, AspectRatio, NumLoops
  - …


0


1

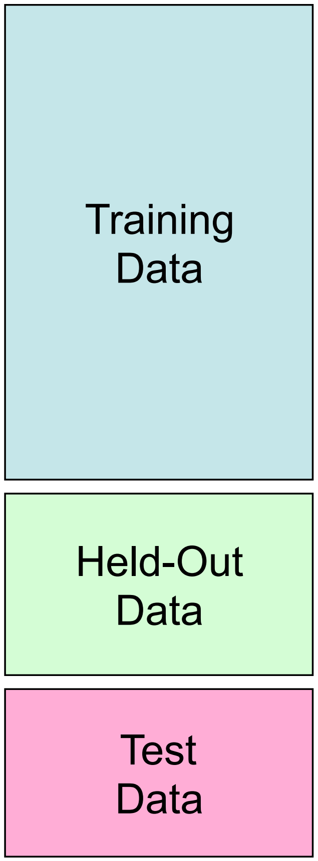

2


1


??

Important Concepts

- Data: labeled instances, e.g. emails marked spam/ham
  - Training set
  - Held out set (sometimes call Validation set)
  - Test set
- Features: attribute-value pairs which characterize each x
- Experimentation cycle
  - Select a hypothesis f to best match training set
  - (Tune hyperparameters on held-out set)
  - Compute accuracy of test set
  - Very important: never "peek" at the test set!
- Evaluation
  - Accuracy: fraction of instances predicted correctly
- Overfitting and generalization
  - Want a classifier which does well on *test* data
  - Overfitting: fitting the training data very closely, but not generalizing well
  - We'll investigate overfitting and generalization formally in a few lectures



A Supervised Learning Problem

- Consider a simple, Boolean dataset:
  - $f : X \rightarrow Y$
  - $X = \{0,1\}^4$
  - $Y = \{0,1\}$
- Question 1: How should we pick the *hypothesis space*, the set of possible functions f?
- Question 2: How do we find the best f in the hypothesis space?

Dataset:

Example	x_1	x_2	x_3	x_4	y
1	0	0	1	0	0
2	0	1	0	0	0
3	0	0	1	1	1
4	1	0	0	1	1
5	0	1	1	0	0
6	1	1	0	0	0
7	0	1	0	1	0

Most General Hypothesis Space

Consider all possible boolean functions over four input features!

- 2^{16} possible hypotheses
- 2^{9} are consistent with our dataset
- How do we choose the best one?

x_1	x_2	x_3	x_4	y
0	0	0	0	?
0	0	0	1	?
0	0	1	0	0
0	0	1	1	1
0	1	0	0	0
0	1	0	1	0
0	1	1	0	0
0	1	1	1	?
1	0	0	0	?
1	0	0	1	1
1	0	1	0	?
1	0	1	1	?
1	1	0	0	0
1	1	0	1	?
1	1	1	0	?
1	1	1	1	?

Dataset:

Example	x_1	x_2	x_3	x_4	y
1	0	0	1	0	0
2	0	1	0	0	0
3	0	0	1	1	1
4	1	0	0	1	1
5	0	1	1	0	0
6	1	1	0	0	0
7	0	1	0	1	0

A Restricted Hypothesis Space

Consider all conjunctive boolean functions.

- 16 possible hypotheses
- None are consistent with our dataset
- How do we choose the best one?

Rule	Counterexample
$\Rightarrow y$	1
$x_1 \Rightarrow y$	3
$x_2 \Rightarrow y$	2
$x_3 \Rightarrow y$	1
$x_4 \Rightarrow y$	7
$x_1 \wedge x_2 \Rightarrow y$	3
$x_1 \wedge x_3 \Rightarrow y$	3
$x_1 \wedge x_4 \Rightarrow y$	3
$x_2 \wedge x_3 \Rightarrow y$	3
$x_2 \wedge x_4 \Rightarrow y$	3
$x_3 \wedge x_4 \Rightarrow y$	4
$x_1 \wedge x_2 \wedge x_3 \Rightarrow y$	3
$x_1 \wedge x_2 \wedge x_4 \Rightarrow y$	3
$x_1 \wedge x_3 \wedge x_4 \Rightarrow y$	3
$x_2 \wedge x_3 \wedge x_4 \Rightarrow y$	3
$x_1 \wedge x_2 \wedge x_3 \wedge x_4 \Rightarrow y$	3

Dataset:

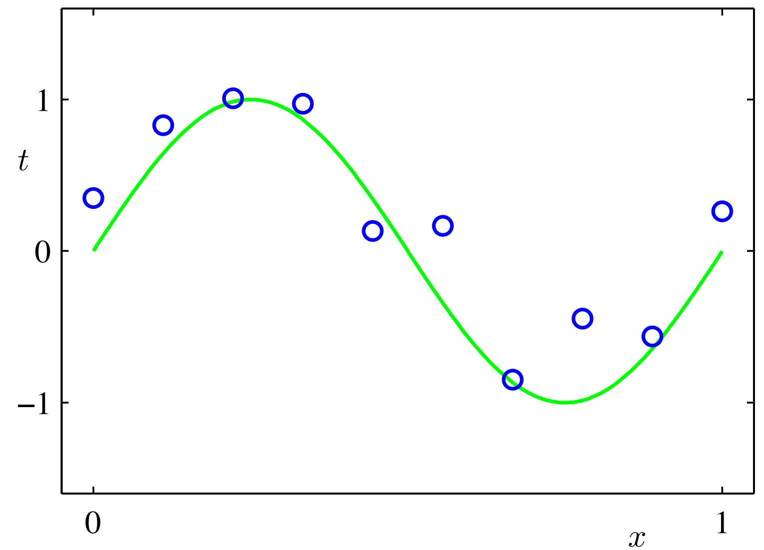
Example	x_1	x_2	x_3	x_4	y
1	0	0	1	0	0
2	0	1	0	0	0
3	0	0	1	1	1
4	1	0	0	1	1
5	0	1	1	0	0
6	1	1	0	0	0
7	0	1	0	1	0

Another Sup. Learning Problem

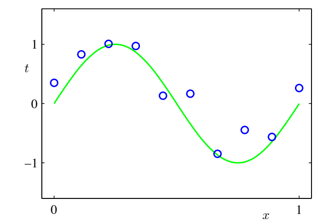
- Consider a simple, regression dataset:
  - $f : X \rightarrow Y$
  - $X = \Re$
  - $Y = \Re$

- Question 1: How should we pick the *hypothesis space*, the set of possible functions f?
- Question 2: How do we find the best f in the hypothesis space?

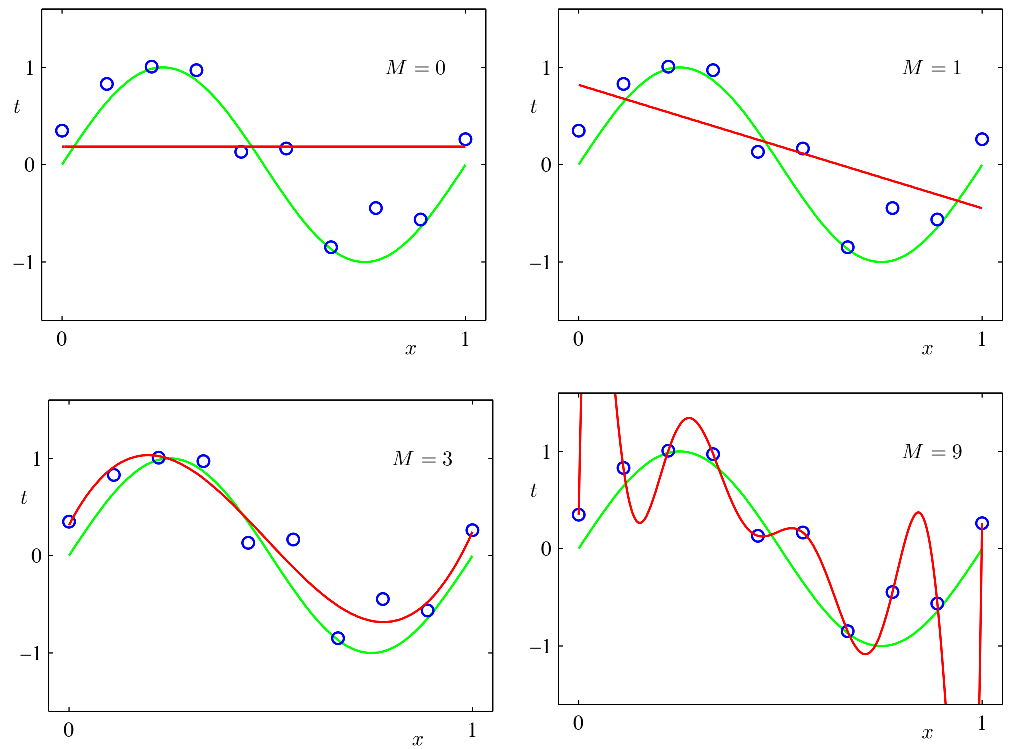
Dataset: 10 points generated from a sin function, with noise

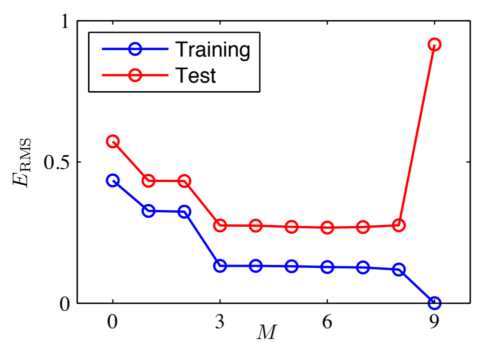


Hypo. Space: Degree-N Polynomials



- Infinitely many hypotheses
- None / Infinitely many are consistent with our dataset
- How do we choose the best one?





Key Issues in Machine Learning

- What are good hypothesis spaces?
- How to find the best hypothesis? (algorithms / complexity)
- How to optimize for accuracy of unseen testing data? (avoid overfitting, etc.)
- Can we have confidence in results? How much data is needed?
- How to model applications as machine learning problems? (engineering challenge)