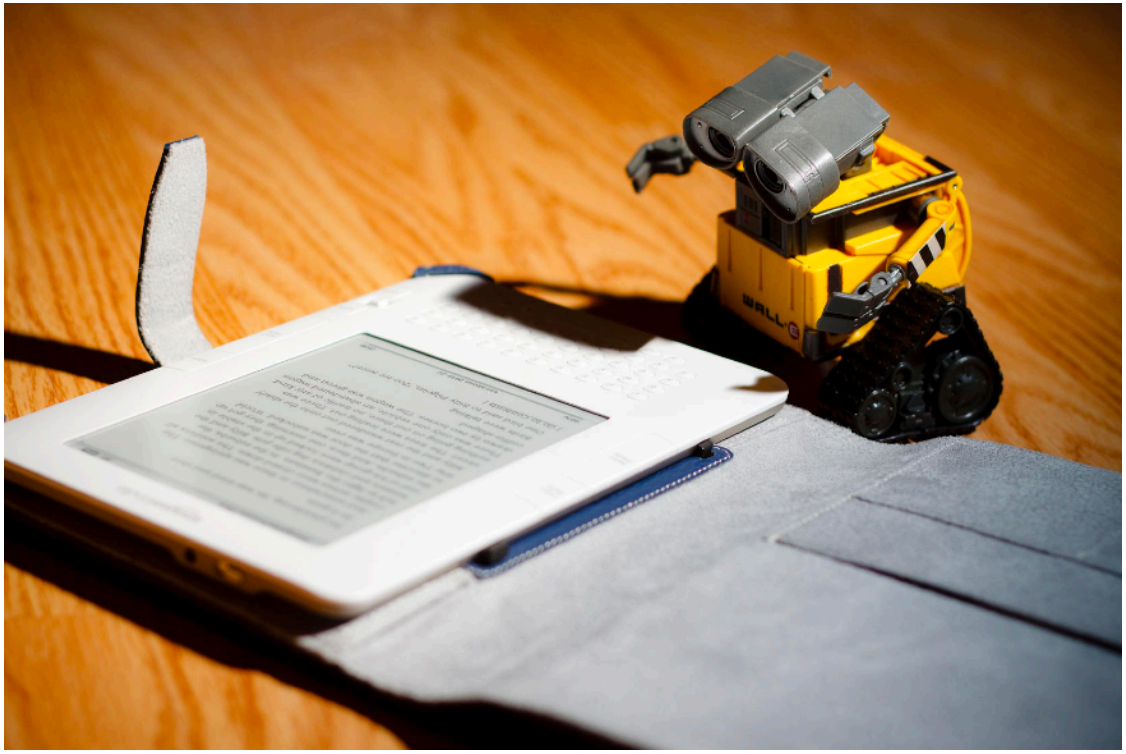
CSE 4641: Machine Learning

Instructor: Alan Ritter



Administrative Details

- **Instructor**
  - Alan Ritter
- Course Webpage
  - **https://aritter.github.io/CS-4641/**
  - Please read through this carefully
- Lectures are being broadcast and recorded using BlueJeans Events
  - **https://primetime.bluejeans.com/a2m/live-event/dcwjksqr**
- All graded assignments will be submitted using Gradescope:
  - https://www.gradescope.com/courses/281746

COVID-19

The New York Times

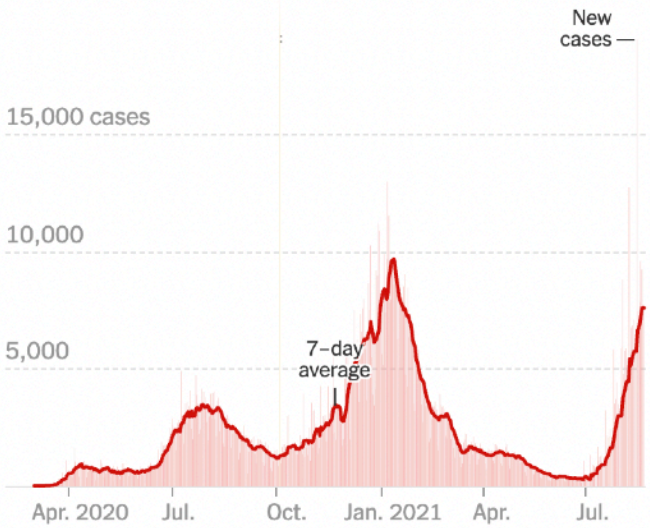
Georgia

	ALL AGES	18 AND UP	65 AND UP
At least one dose	49%	61%	84%
Fully vaccinated	40%	50%	73%

See more details ›

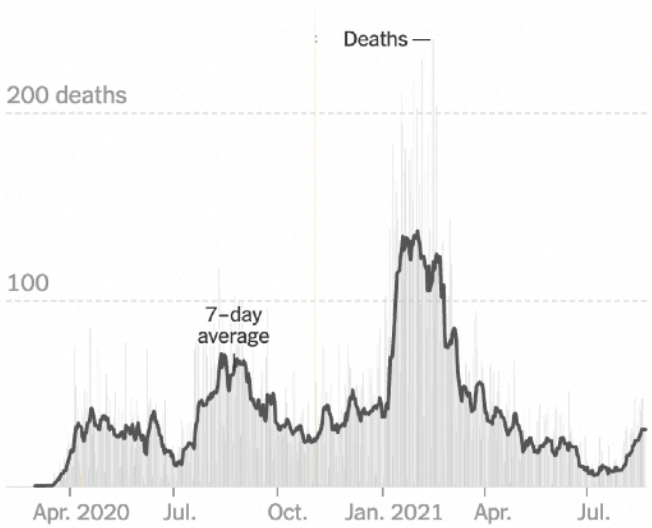
About this data

New cases



These are days with a data reporting anomaly. Read more here.

New deaths



CDC Guidance

https://www.cdc.gov/coronavirus/2019-ncov/vaccines/fully-vaccinated-guidance.html

Infections happen in only a small proportion of people who are fully vaccinated, even with the Delta variant. However, preliminary evidence suggests that fully vaccinated people who do become infected with the Delta variant can spread the virus to others. To reduce their risk of becoming infected with the Delta variant and potentially spreading it to others: CDC recommends that fully vaccinated people:

- Wear a mask in public indoor settings if they are in an area of substantial or high transmission.
  - Fully vaccinated people might choose to mask regardless of the level of transmission, particularly if they or someone in their household is immunocompromised or at increased risk for severe disease, or if someone in their household is unvaccinated. People who are at increased risk for severe disease include older adults and those who have certain medical conditions, such as diabetes, overweight or obesity, and heart conditions.

CDC Guidance

https://www.cdc.gov/coronavirus/2019-ncov/vaccines/fully-vaccinated-guidance.html

Infections happen in only a small proportion of people who are fully vaccinated, even with the Delta variant. However, preliminary evidence suggests that fully vaccinated people who do become infected with the Delta variant can spread the virus to others. To reduce their risk of becoming infected with the Delta variant and potentially spreading it to others: CDC recommends that fully vaccinated people:

- Wear a mask in public indoor settings if they are in an area of substantial or high transmission.
  - Fully vaccinated people might choose to mask regardless of the level of transmission, particularly if they or someone in their household is immunocompromised or at increased risk for severe disease, or if someone in their household is unvaccinated. People who are at increased risk for severe disease include older adults and those who have certain medical conditions, such as diabetes, overweight or obesity, and heart conditions.

CDC Guidance

https://www.cdc.gov/coronavirus/2019-ncov/vaccines/fully-vaccinated-guidance.html

Infections happen in only a small proportion of people who are fully vaccinated, even with the Delta variant. However, preliminary evidence suggests that fully vaccinated people who do become infected with the Delta variant can spread the virus to others. To reduce their risk of becoming infected with the Delta variant and potentially spreading it to others: CDC recommends that fully vaccinated people:

- Wear a mask in public indoor settings if they are in an area of substantial or high transmission.
  - Fully vaccinated people might choose to mask regardless of the level of transmission, particularly if they or someone in their household is immunocompromised or at increased risk for severe disease, or if someone in their household is unvaccinated. People who are at increased risk for severe disease include older adults and those who have certain medical conditions, such as diabetes, overweight or obesity, and heart conditions.

Level of Community Transmission in Fulton County, Georgia

CDC Guidance

https://www.cdc.gov/coronavirus/2019-ncov/vaccines/fully-vaccinated-guidance.html

Infections happen in only a small proportion of people who are fully vaccinated, even with the Delta variant. However, preliminary evidence suggests that fully vaccinated people who do become infected with the Delta variant can spread the virus to others. To reduce their risk of becoming infected with the Delta variant and potentially spreading it to others: CDC recommends that fully vaccinated people:

- Wear a mask in public indoor settings if they are in an area of substantial or high transmission.
  - Fully vaccinated people might choose to mask regardless of the level of transmission, particularly if they or someone in their household is immunocompromised or at increased risk for severe disease, or if someone in their household is unvaccinated. People who are at increased risk for severe disease include older adults and those who have certain medical conditions, such as diabetes, overweight or obesity, and heart conditions.

Level of Community Transmission in Fulton County, Georgia

Please wear a mask while you are in this class!

Fall 2021
Tech Moving Forward

For Our Safety

- Masks/face coverings are recommended while inside campus facilities.
- All faculty, staff, and students are strongly encouraged to receive COVID-19 vaccines.
- Stamps Health Services is offering free Covid-19 vaccines at McCamish Pavilion in August and September. https://health.gatech.edu/coronavirus/vaccine
  - To make an appointment, go to https://mytest.gatech.edu
- Asymptomatic testing on campus is easy, convenient and free: https://health.gatech.edu/coronavirus/testing
- For updates, guidelines and Q&As: https://health.gatech.edu/tech-moving-forward

Georgia Tech
CREATING THE NEXT

From our President. . . .

Ángel Cabrera
@CabreraAngel

It's on each of us @georgiatech to protect one another, stay healthy + have a thriving campus:
- Please vaccinate if you haven't already
- Please wear your mask in class, labs and buildings where we can't keep appropriate distance
- Please test regularly
president.gatech.edu/blog/jackets-m...

2:34 PM · Aug 16, 2021 · Twitter Web App

Georgia Tech
CREATING THE NEXT

Office Hours

- TA Office Hours Spreadsheet:
  - https://docs.google.com/spreadsheets/d/17Nkv86GstnlVbKiEcqLMuX-bPEcfVY4k2xnG2IFf0qA/edit#gid=0
  - Add your name to the list before the start, otherwise the TA will assume nobody is attending office hours.

Administrative Details

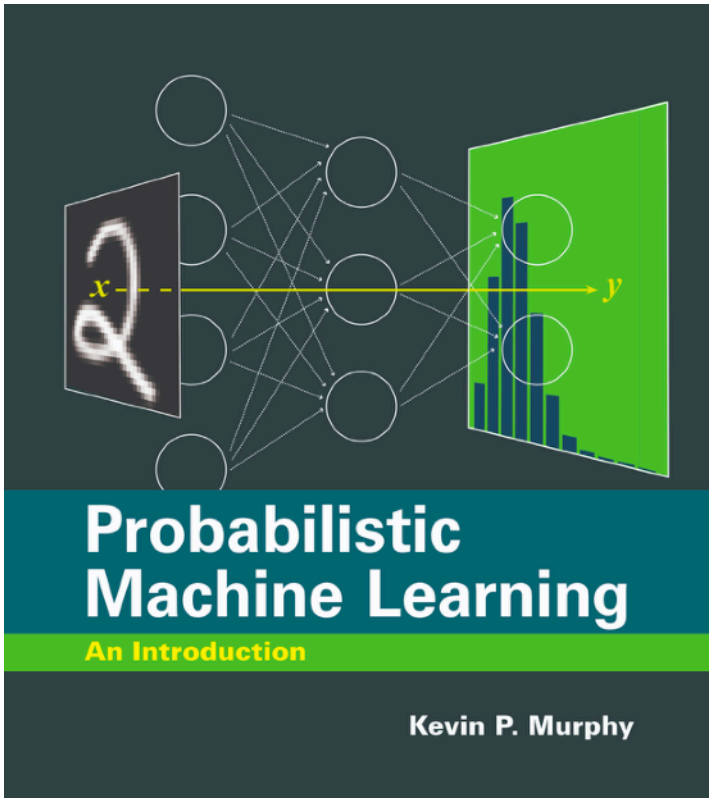
- **Piazza** (discussion and resources):
  - https://piazza.com/class/krjfpfjr3es38i
  - Best place to ask questions about anything related to the class.
  - Please ask publicly, if the answer may be helpful for other students
  - We will do our best to answer your questions within 24 hours (keep in mind TAs might not be available to answer questions over the weekend).
- **GradeScope** (homework submissions)
  - https://www.gradescope.com/courses/281746

Administrative Details

- **Waitlisted Students**
  - Last I checked there were 66 students on the waitlist.
  - I don’t know how likely it is any specific students will be able to get in based on their waitlist position, etc.
  - I am unable to sign any forms to circumvent the waitlist / course prerequisites, etc.
  - Please consult with the academic advisors in your department/school for more information.

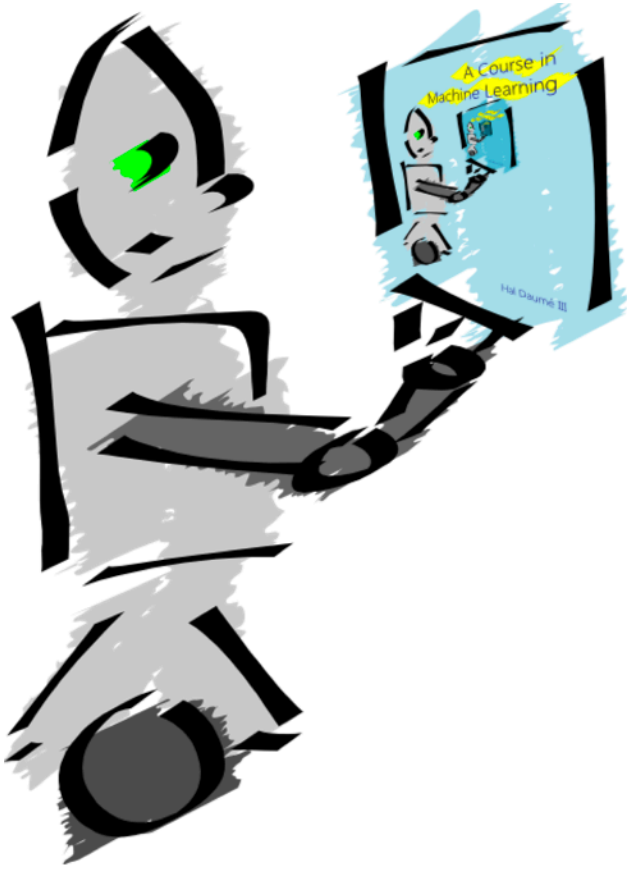
Reading: 2 Books

- Probabilistic Machine Learning:
  - https://probml.github.io/pml-book/book1.html



Reading: 2 Books

- H. Daume III, A Course in Machine Learning
  - http://ciml.info/



Prerequisites

- This class assumes that you know:
  - Probability
  - Linear Algebra
  - Calculus
  - Python (or have ability to learn Python quickly)
  - Numpy/Scipy Libraries (or can learn them quickly).

Prerequisites

- This class assumes that you know:
  - Probability
  - Linear Algebra
  - Calculus
  - Python (or have ability to learn Python quickly)
  - Numpy/Scipy Libraries (or can learn them quickly).

- **Homework #1 is out**
  - Due next week
  - Turn in on GradeScope

Grading (from course website)

Programming Assignments (Projects) - 40%

We plan to assign three or four programming assignments that provide hands-on experience implementing algorithms discussed during lecture. The assignments are in Python. Completing these projects will often require waiting for your models to train (this can range up to about 30 minutes to hours depending on the efficiency of your implementation), so we strongly recommend starting work on these programming assignments well in advance of the deadline, as ML algorithms are more difficult to debug than traditional computer programs. If you start working on an assignment the day before it is due, it is unlikely you will be able to complete it on time.

Written Assignments (Problem Sets) - 20%

The written assignments will mostly be mathematical. You can scan and upload your solution to Gradescope. Please write answers clearly, since we won't be able to provide credit for answers that we are not legible.

Midterm Exam - 15%

The midterm will be similar in format to the written assignments (problem sets), but will be more substantial.

Participation - 5%

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

Final Project - 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 2-4. It is a good idea to discuss your planned project with the instructor to get feedback. The final project report should be 4 pages. The report should describe the problem you are solving, what data is being used, the proposed technique you are applying in addition to what baseline is used to compare against.

What to Expect

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 programing assignments early.
    - (In general, these won't be possible to complete if you wait to start until the day before)

- Questions?

A Few Quotes

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

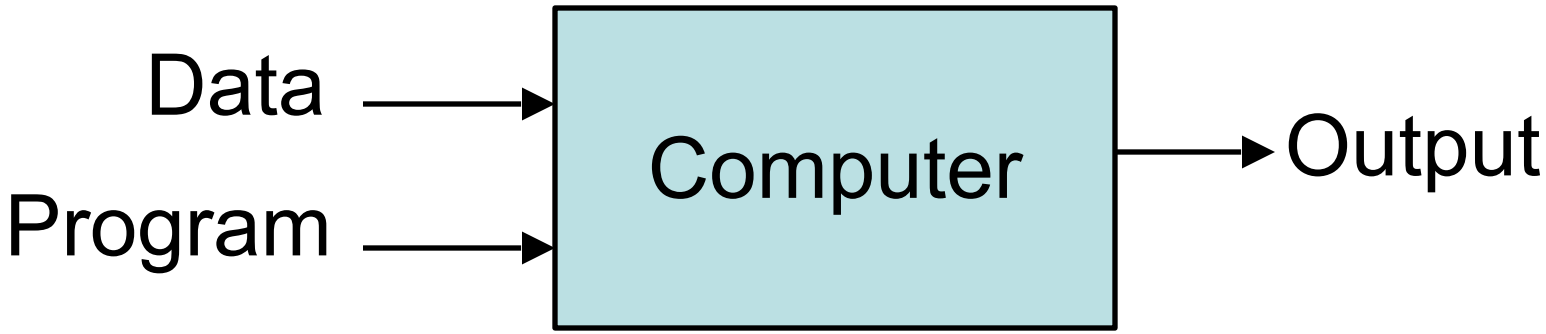
A Few Quotes

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

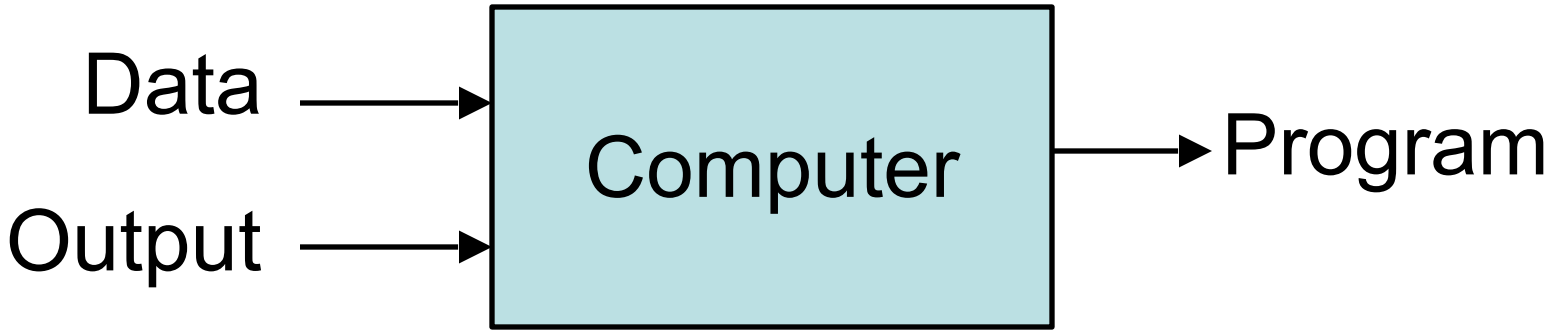
So What Is Machine Learning?

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

Traditional Programming



Machine Learning



Magic?

Magic?

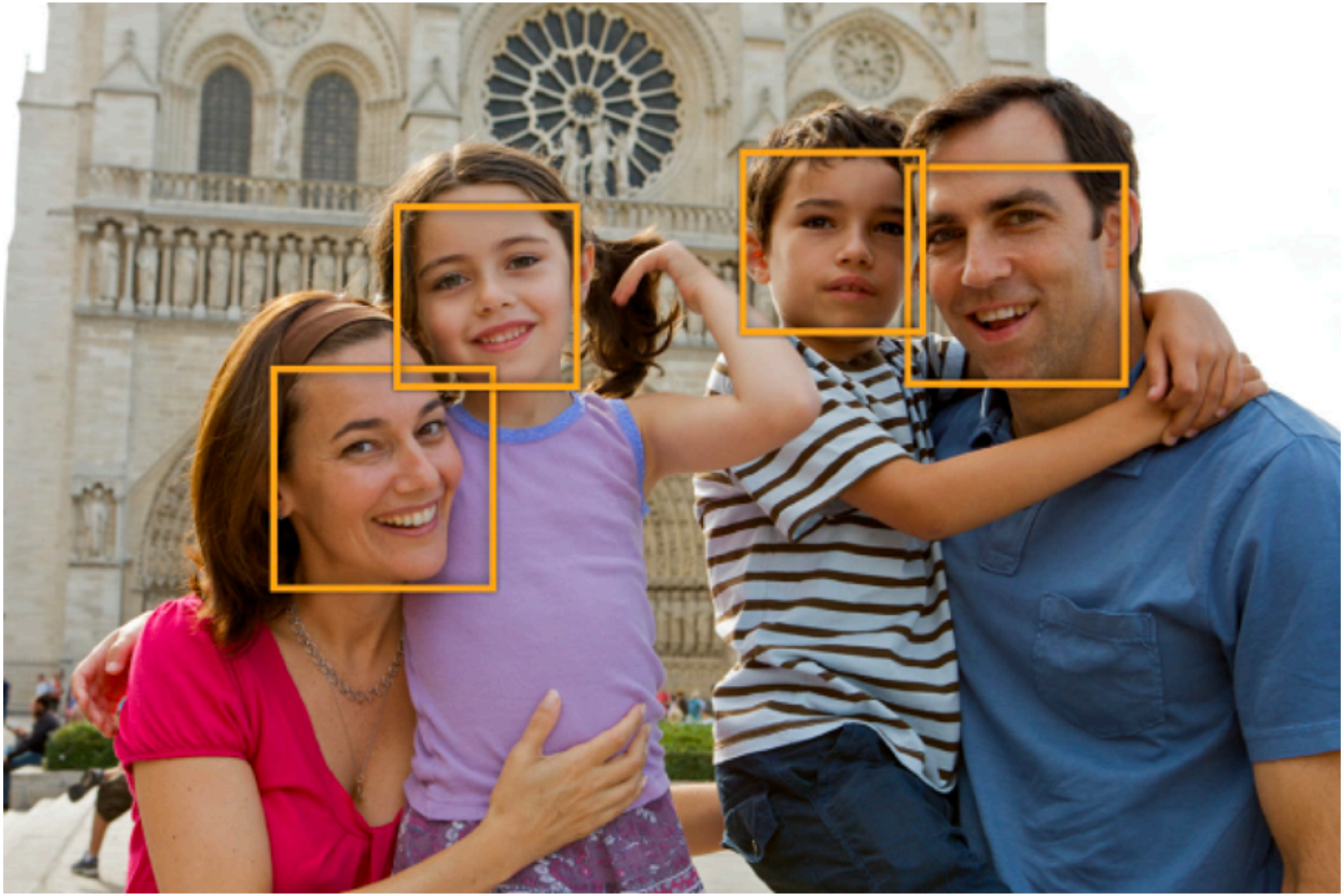
No, more like farming

Magic?

No, more like farming

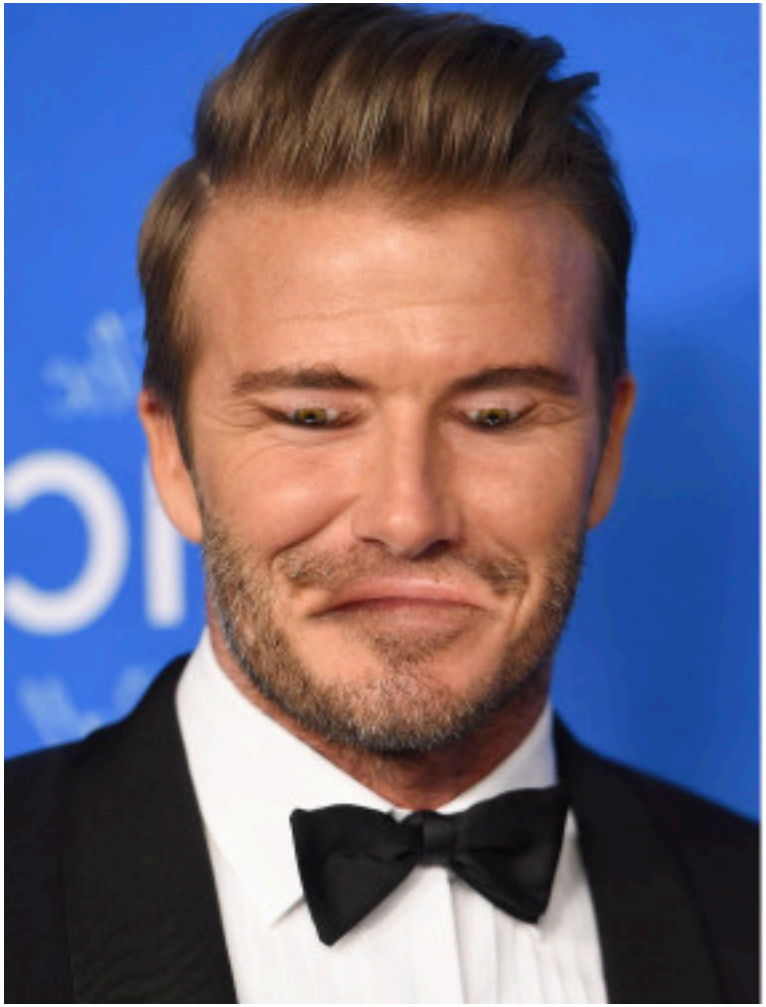
- **Seeds** = Learning Algorithms
- **Nutrients** = Data
- **Farmer** = You
- **Plants** = Programs

Sample Applications



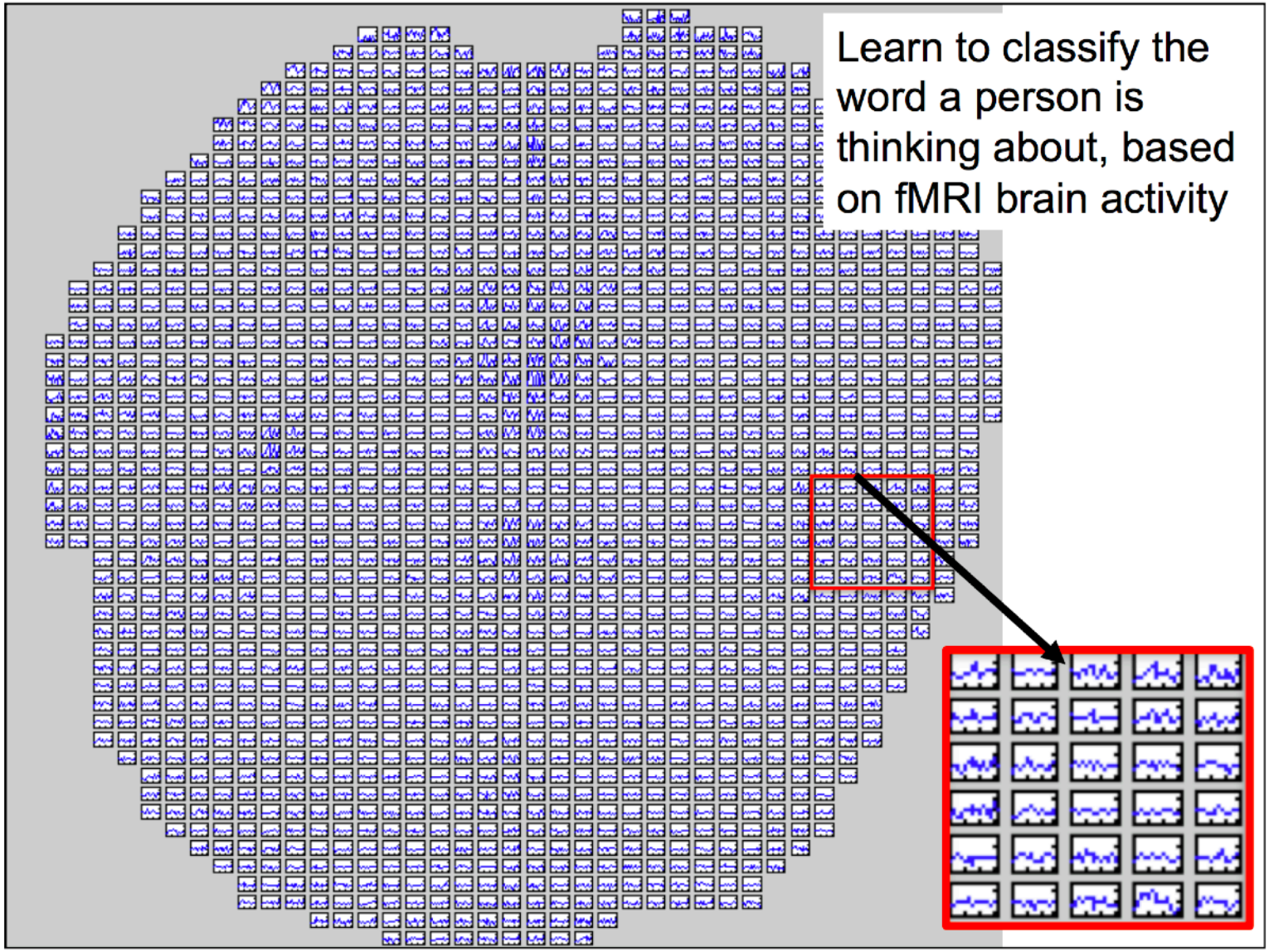
Sample Applications

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

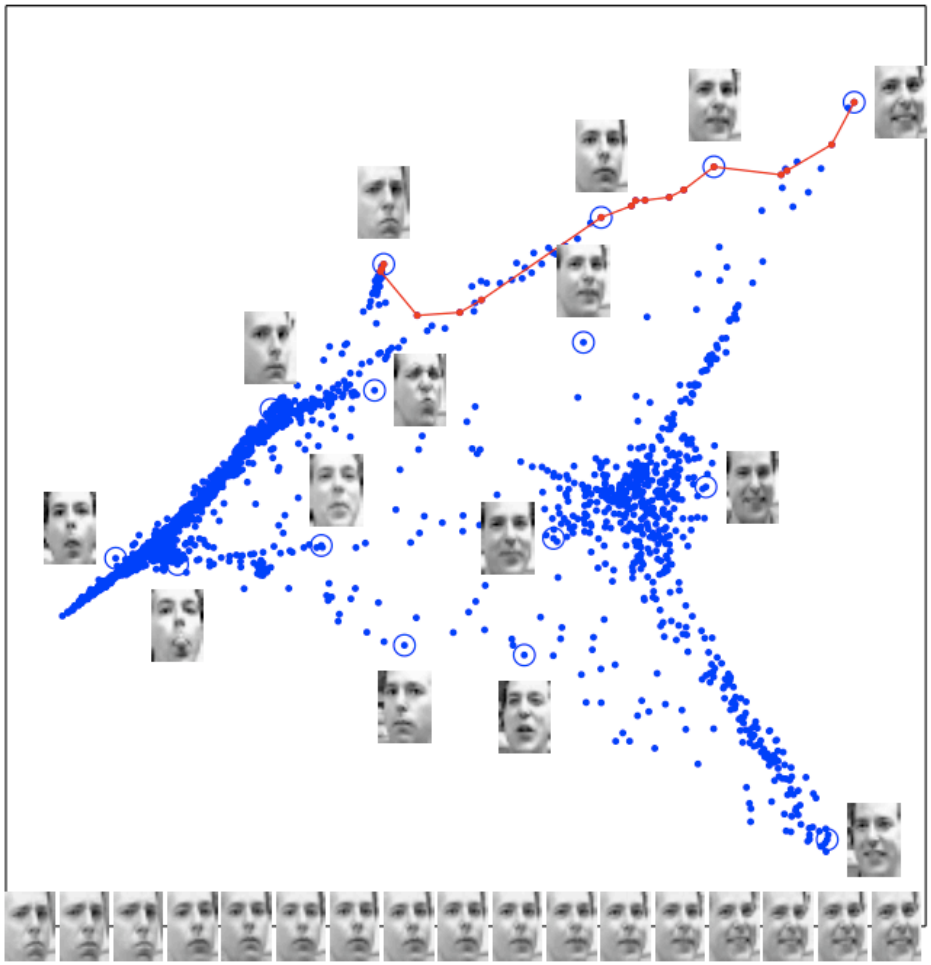
- Reinforcement skill learning
- Semi-supervised learning
- Active student querying
- …

… also relating:

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

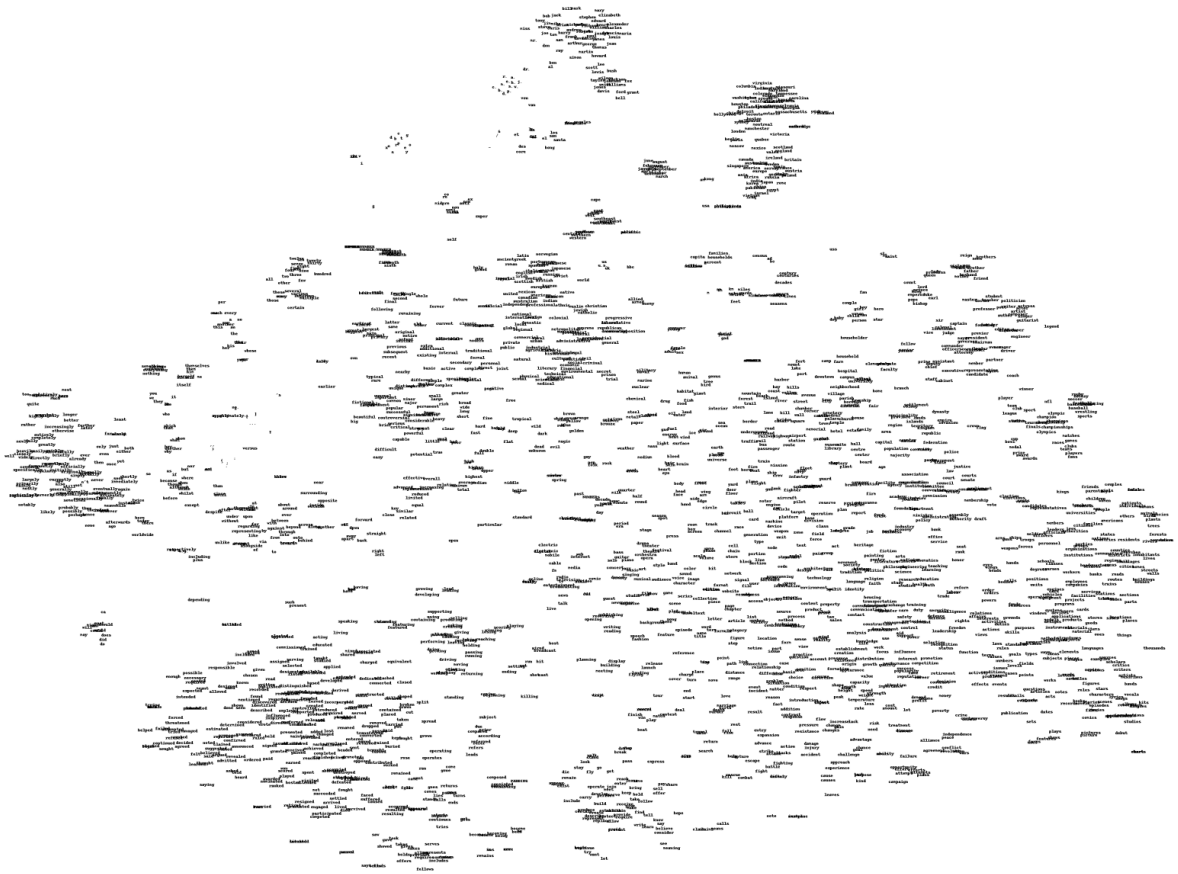
Embedding images

- 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

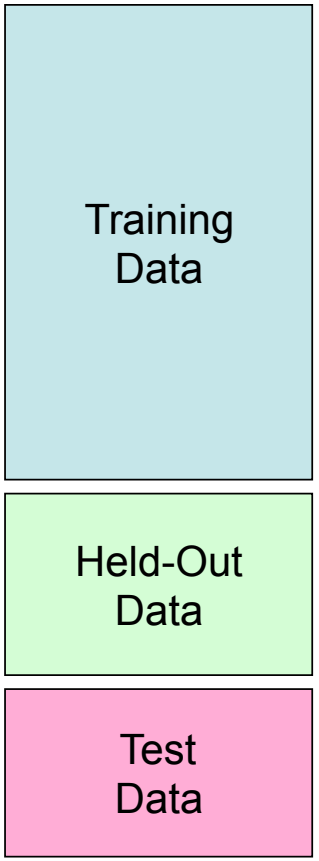
2

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:

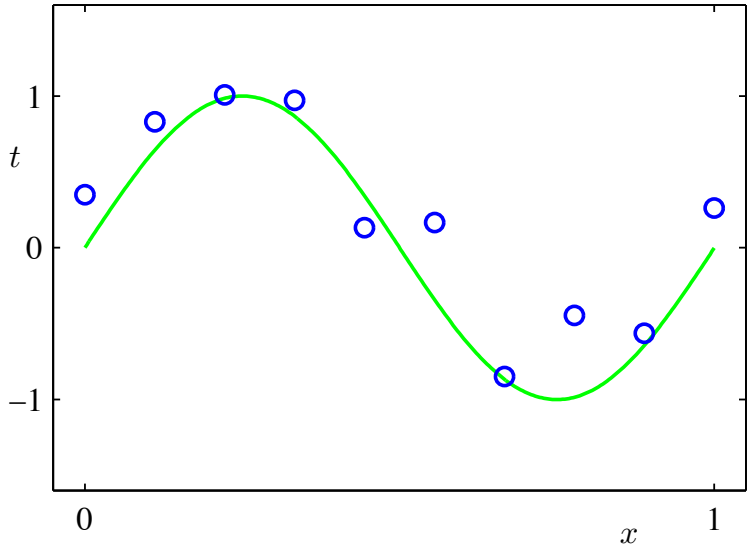
Example	x_1	x_2	x_3	x_4	y
1	0	0	1	0	0
2	0	1	0	0	0
3	0	0	1	1	1
4	1	0	0	1	1
5	0	1	1	0	0
6	1	1	0	0	0
7	0	1	0	1	0

Another Sup. Learning Problem

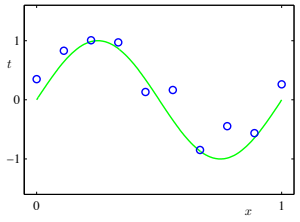
- Consider a simple, regression dataset:
  - $f : X \rightarrow Y$
  - $X = \Re$
  - $Y = \Re$

- 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?

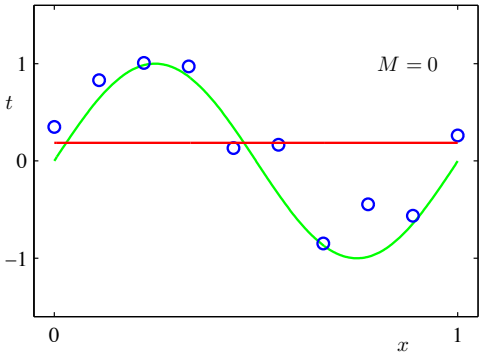
Dataset: 10 points generated from a sin function, with noise

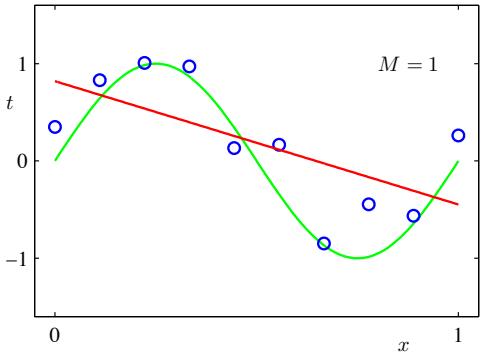


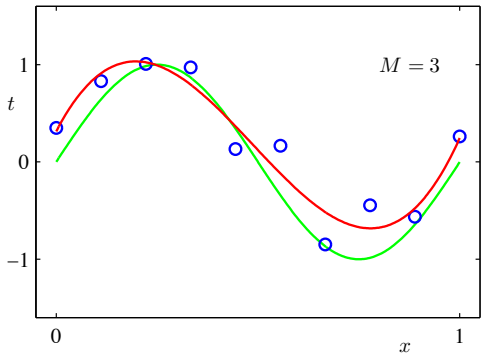
Hypo. Space: Degree-N Polynomials

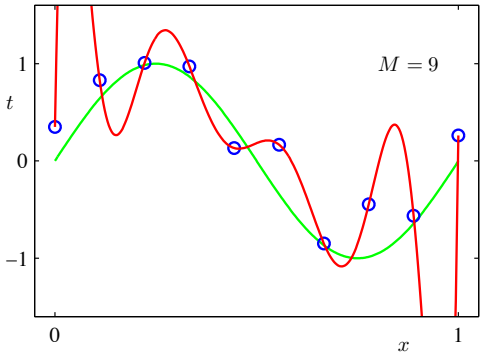


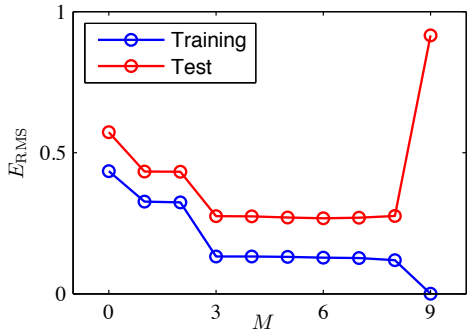
- Infinitely many hypotheses
- None / Infinitely many are consistent with our dataset
- How do we choose the best one?











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)