CSE 4/574 Introduction to Machine Learning

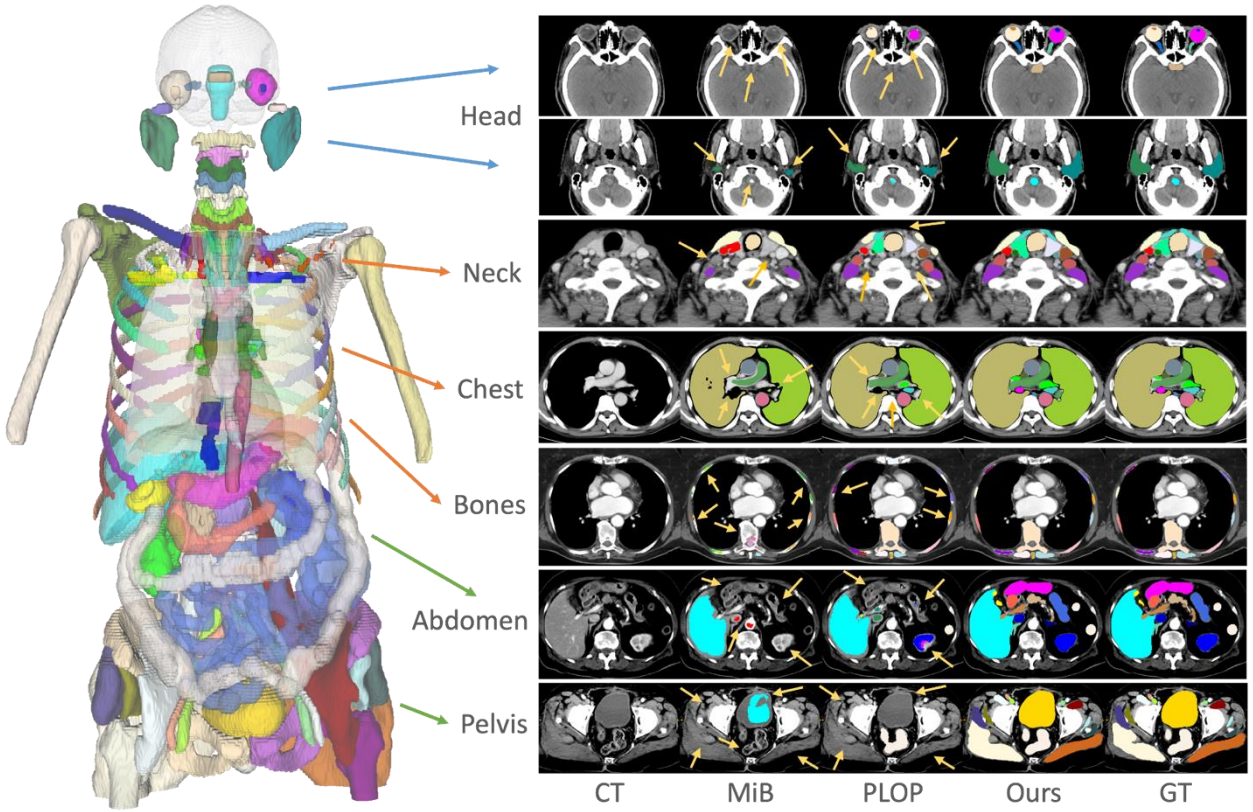
Mingchen Gao
mgao8@buffalo.edu
Davis ~~347~~317

Slides adapted from Varun Chandola and Luke Zettlemoyer

Logistics

- Class webpage
- https://piazza.com/buffalo/fall2024/cse474574/resources
- Office hours (Davis 317)
  - Tuesdays 2 – 3pm
- TAs
  - Graduate TAs
    - Shaoshu Su, shaoshus@buffalo.edu
    - Chao Wu, cwu64@buffalo.edu
    - Kangxian Xie, kangxian@buffalo.ede
    - Pouya Karimian,

Who am I? Research and Personnel



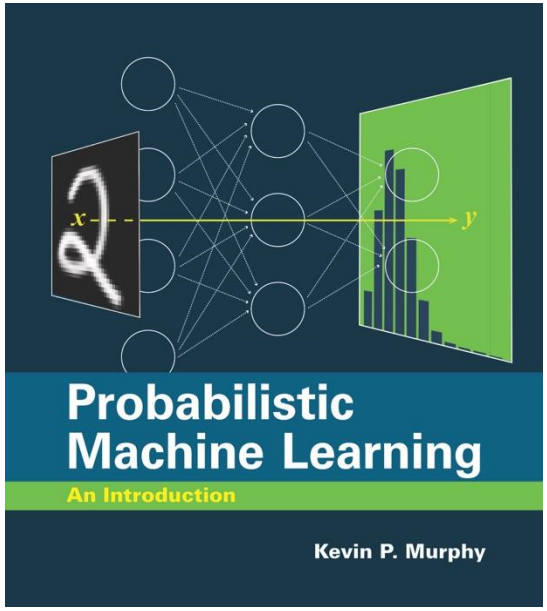
Most of my time in the office is spent doing research, medical imaging analysis
Outside of office, I spent a lot of time with my little one

Topics Covered

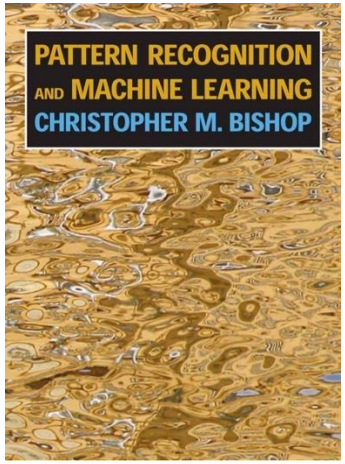
- Machine Learning Tools:
  - Bayesian Learning
  - Expectation Maximization
  - Optimization
- Machine Learning Algorithms:
  - Generative Models
  - Regression
  - Logistic Regression
  - Perceptron and Neural Networks
  - Convolutional Neural Networks, Recurrent Neural Networks
  - Graphical Models
  - Latent Linear Models
  - Support Vector Machines

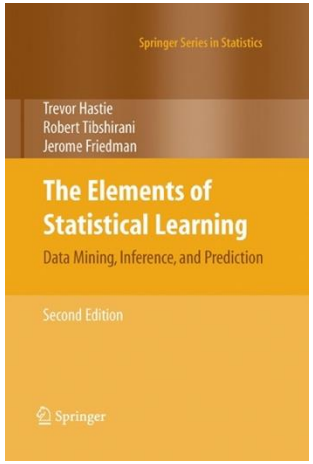
Textbooks

- No textbook is required.
- Kevin Murphy, Probabilistic Machine Learning, MIT Press, 2022. https://probml.github.io/pml-book/book1.html



- Reference textbooks:
  - Christopher Bishop, Pattern Recognition and Machine Learning, Springer, 2007
  - Tom Mitchell, Machine Learning, McGraw-Hill, 1997
  - Trevor Hastie, Robert Tibshirani and Jerome Friedman, The Elements of Statistical Learning. Springer, 2009.





Prerequisites

- Probability and Statistics
- Linear Algebra
- Python Programming

Evaluation

- Attendance and in-class quizzes – 10%
- Short weekly quizzes on UBlearns (12) – 20%
- Programming Assignments (3) – 30%
  - Group project of up to three students
  - Python, use UBlearns for all electronic submissions
- Two Mid-terms (in-class, open book) – 20%
  - Tentative 10/01/2024 Tue, in class – 10%
  - Tentative 11/05/2024 Tue, in class – 10%
- Final Exam (open book) – 20%
  - 12/16/2024 Mon, 7:15pm – 10:15pm, Knox 104

Final Grade

Graduate

A [92.5; 100]
A- [87.5; 92.5)
B+ [82.5; 87.5)
B [77.5; 82.5)
B- [72.5; 77.5)
C+ [67.5; 72.5)
C [62.5; 67.5)
C- [57.5; 62.5)
D+ [52.5; 57.5)
D [0; 52.5)

Undergraduate

A [87.5; 100]
A- [82.5; 87.5)
B+ [77.5; 82.5)
B [72.5; 77.5)
B- [67.5; 72.5)
C+ [62.5; 67.5)
C [57.5; 62.5)
C- [52.5; 57.5)
D+ [47.5; 52.5)
D [0; 47.5)

Attendance and in-class quizzes (Tophat)

- Instructions will be given on how to signup Tophat, using your ub-email for signup.
- Subscription is needed.

Weekly Quiz

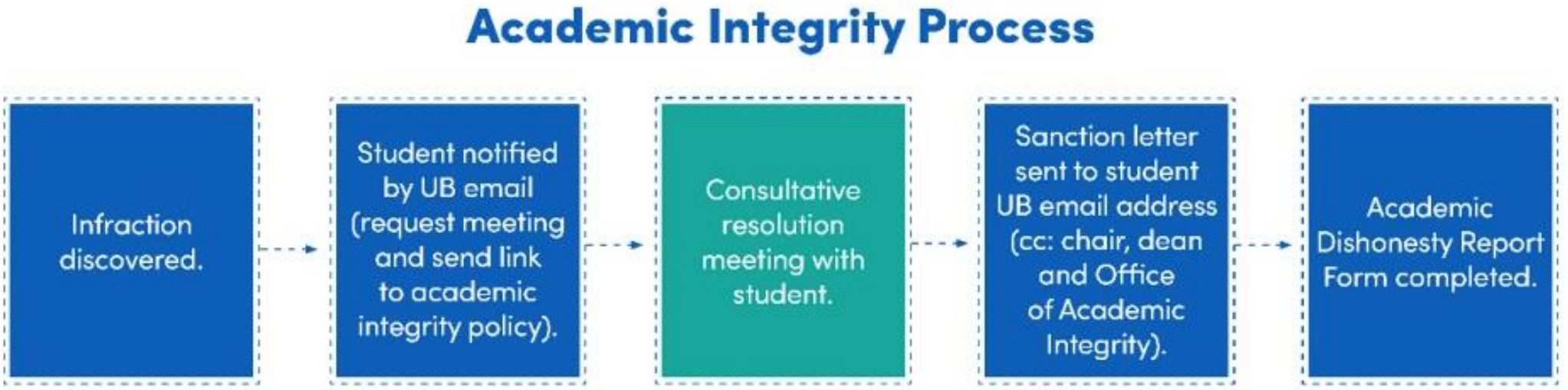
- One quiz per week released on Monday by 8:59 am and due next Sunday by 11:59 pm
- 2-3 multiple choice problems about topics covered that week
- A warm-up quiz (Quiz 0) is posted
- Only 3 tries allowed, maximum score will be used
- Every wrong answer will result in 1 negative point per try

Piazza

- Primary medium of communication
- All announcements, teaching slides, assignments, etc. will be made available through Piazza.
- Questions?
- General post to all (Name will be visible).
- Choose appropriate folder.
- Private post to instructor, TA.
- Interact.

Office of Academic Integrity

https://www.buffalo.edu/academic-integrity/instructors.html



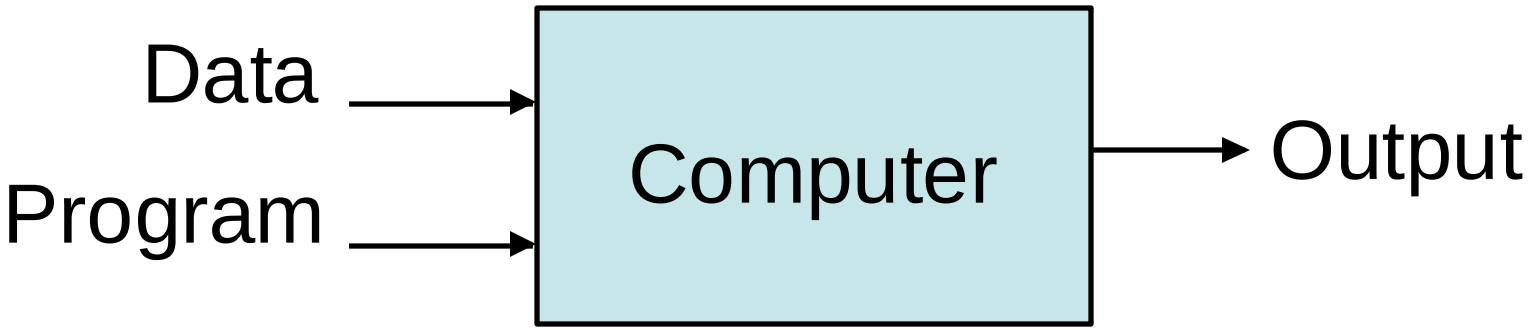
A Few Quotes

- “A breakthrough in machine learning would be worth ten Microsofts” (Bill Gates, Chairman, Microsoft)
- “Machine learning is the next Internet”
  (Tony Tether, Director, DARPA)
- Machine learning is the hot new thing”
  (John Hennessy, President, Stanford)
- “Web rankings today are mostly a matter of machine learning” (Prabhakar Raghavan, Dir. Research, Yahoo)
- “Machine learning is going to result in a real revolution” (Greg Papadopoulos, CTO, Sun)
- “Machine learning is today’s discontinuity”
  (Jerry Yang, CEO, Yahoo)

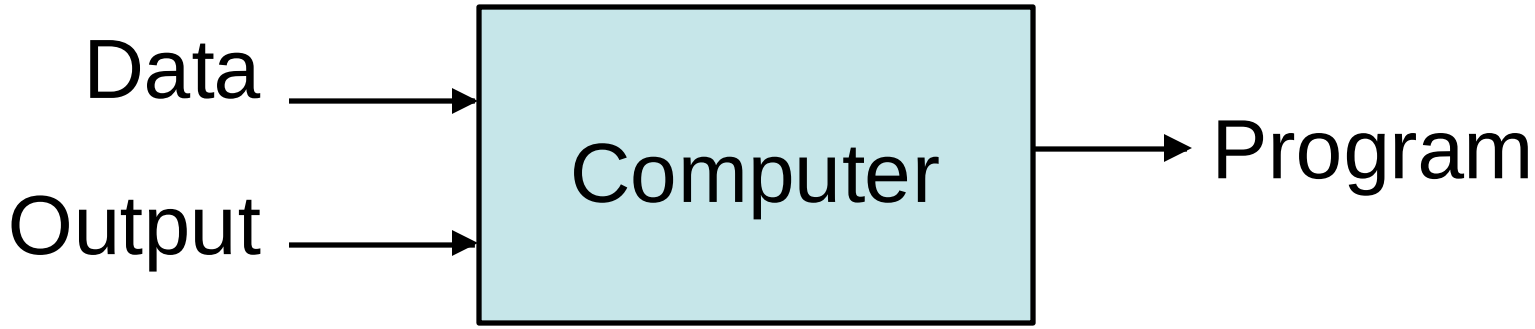
So What Is Machine Learning?

- A set of methods
  - Can automatically detect patterns in data
  - Use uncovered patterns to predict future data under uncertainty
- Let the data do the work instead!
- The future of Computer Science!!!

Traditional Programming



Machine Learning



What is Machine Learning ?
(by examples)

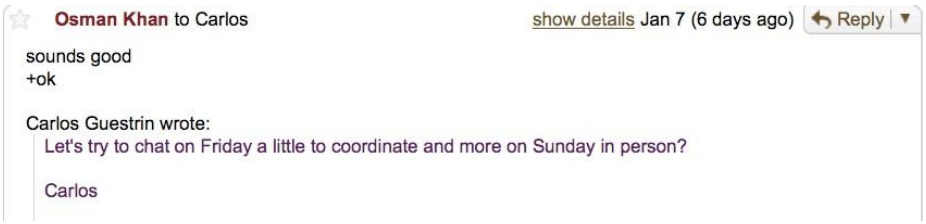
Classification

from data to discrete classes

Spam filtering

data

prediction

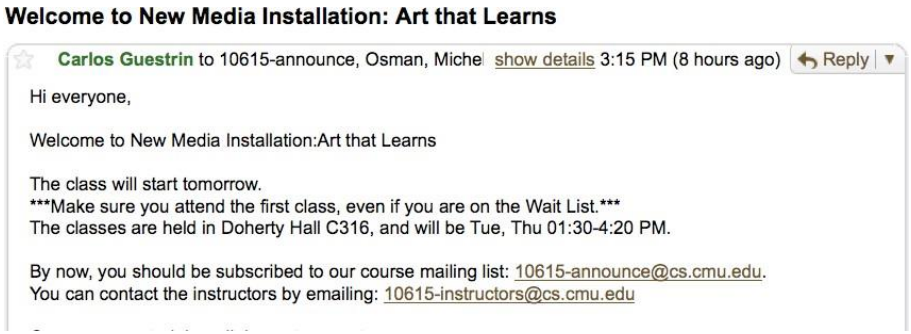


Osman Khan to Carlos show details Jan 7 (6 days ago) Reply

sounds good
+ok

Carlos Guestrin wrote:
Let's try to chat on Friday a little to coordinate and more on Sunday in person?

Carlos



Welcome to New Media Installation: Art that Learns

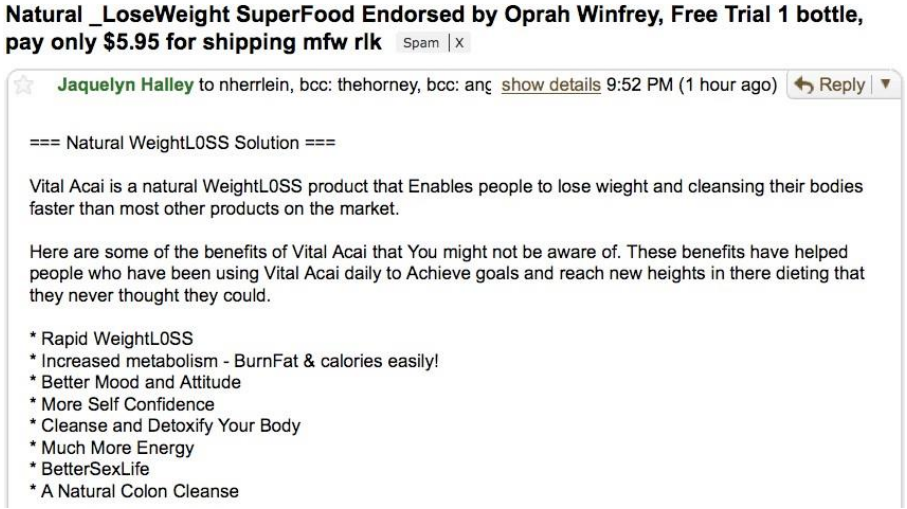
Carlos Guestrin to 10615-announce, Osman, Miche show details 3:15 PM (8 hours ago) Reply

Hi everyone,

Welcome to New Media Installation:Art that Learns

The class will start tomorrow.
Make sure you attend the first class, even if you are on the Wait List.
The classes are held in Doherty Hall C316, and will be Tue, Thu 01:30-4:20 PM.

By now, you should be subscribed to our course mailing list: 10615-announce@cs.cmu.edu.
You can contact the instructors by emailing: 10615-instructors@cs.cmu.edu



Natural _LoseWeight SuperFood Endorsed by Oprah Winfrey, Free Trial 1 bottle, pay only $5.95 for shipping mfw rlk Spam | x

Jaquelyn Halley to nherrlein, bcc: thehorney, bcc: anc show details 9:52 PM (1 hour ago) Reply

=== Natural WeightL0SS Solution ===

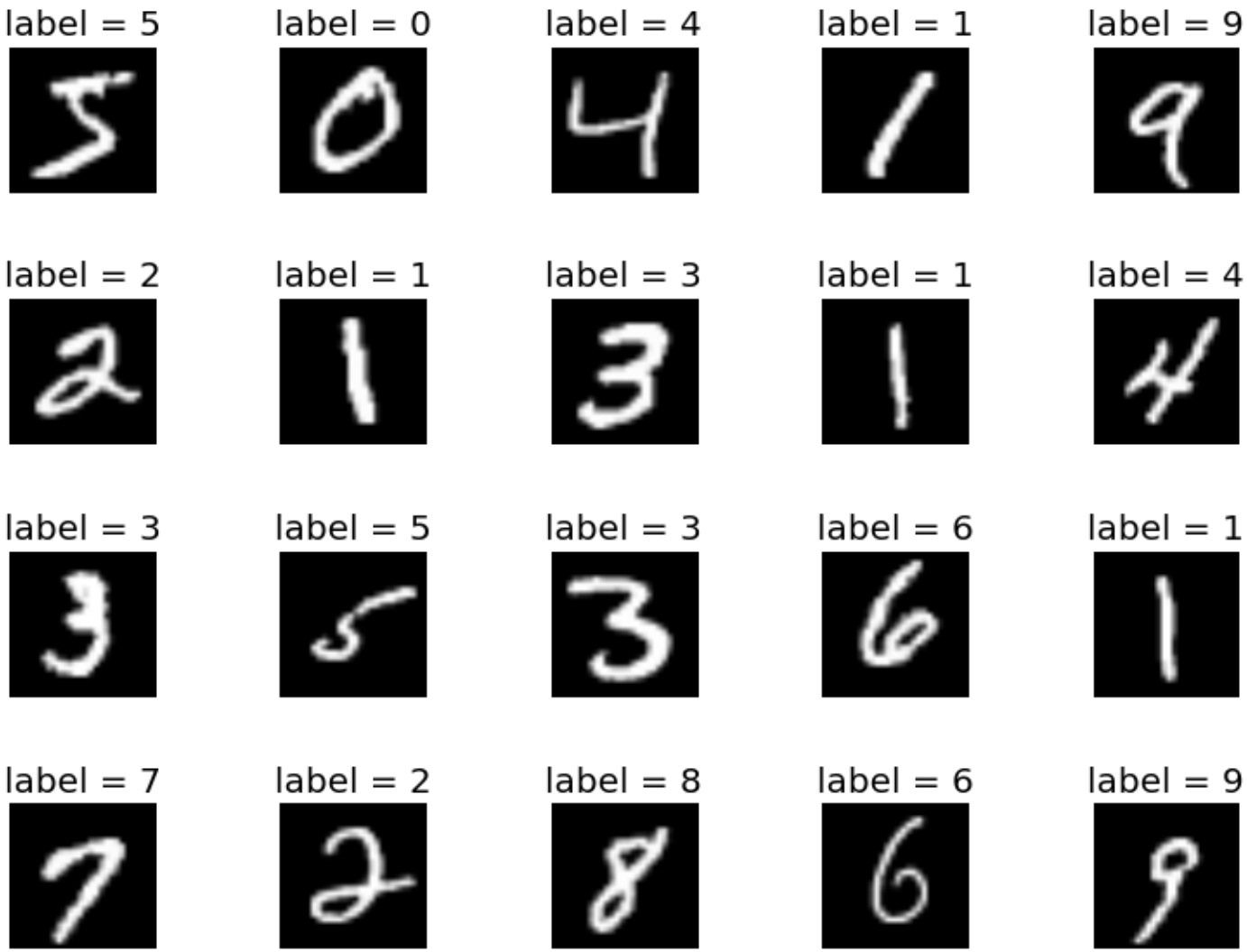
Vital Acai is a natural WeightL0SS product that Enables people to lose wieght and cleansing their bodies faster than most other products on the market.

Here are some of the benefits of Vital Acai that You might not be aware of. These benefits have helped people who have been using Vital Acai daily to Achieve goals and reach new heights in there dieting that they never thought they could.

* Rapid WeightL0SS
* Increased metabolism - BurnFat & calories easily!
* Better Mood and Attitude
* More Self Confidence
* Cleanse and Detoxify Your Body
* Much More Energy
* BetterSexLife
* A Natural Colon Cleanse

Spam
vs
Not Spam

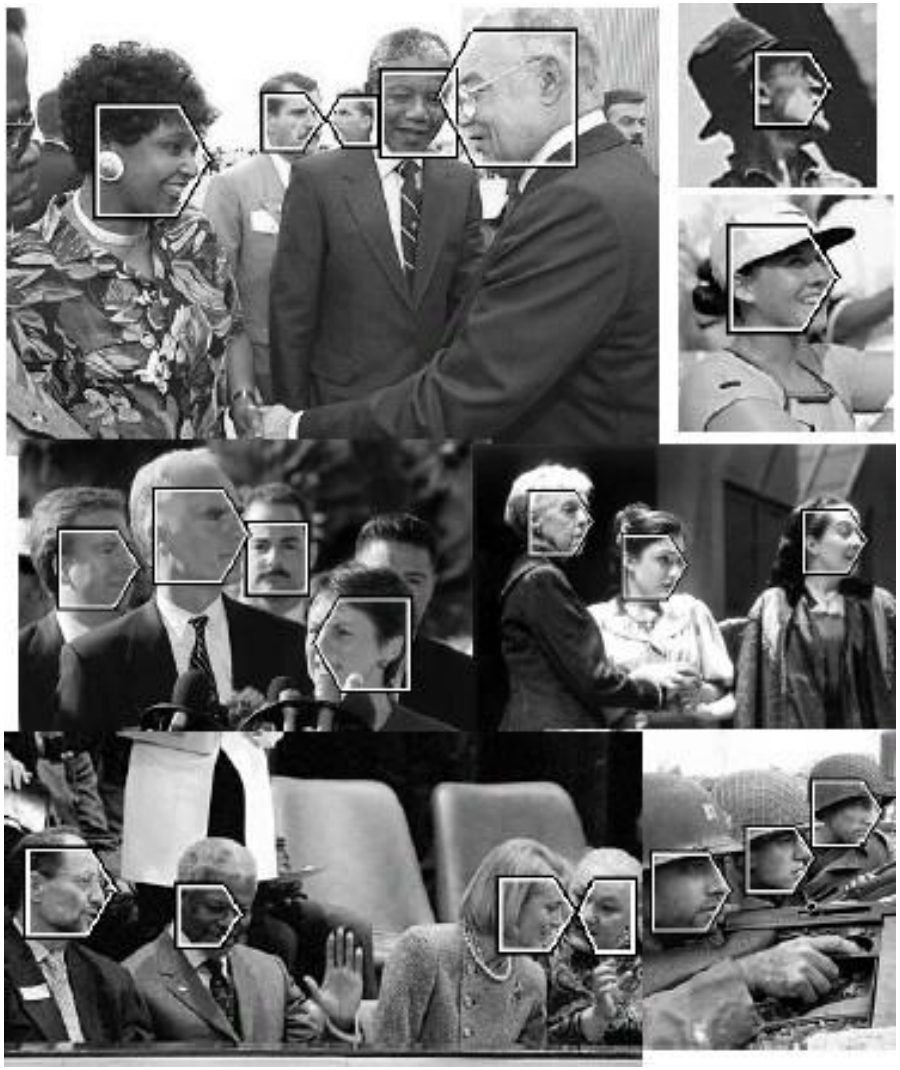
Image classification



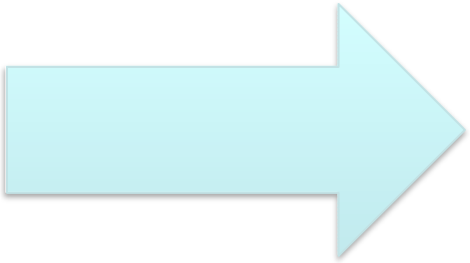
Object detection

(Prof. H. Schneiderman)

Example training images
for each orientation



Weather prediction

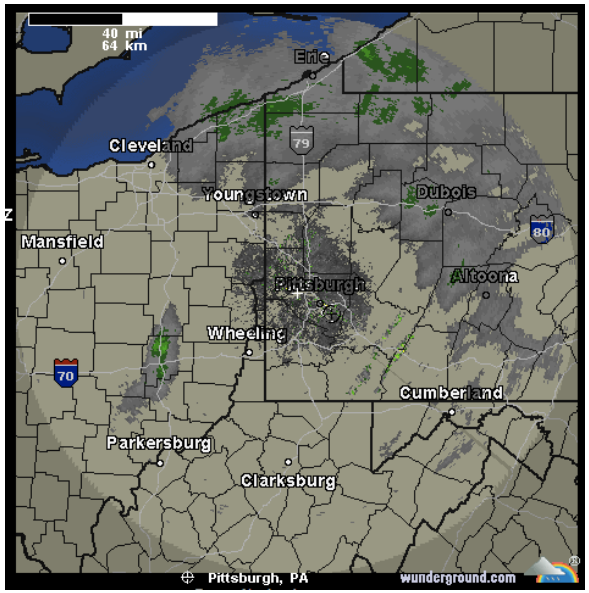


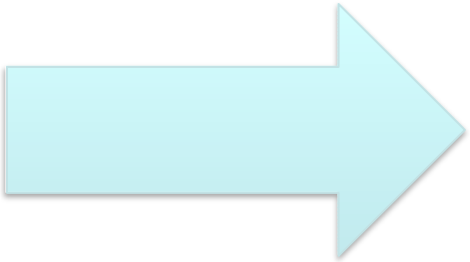
Regression

predicting a numeric value

Stock market

Weather prediction revisted



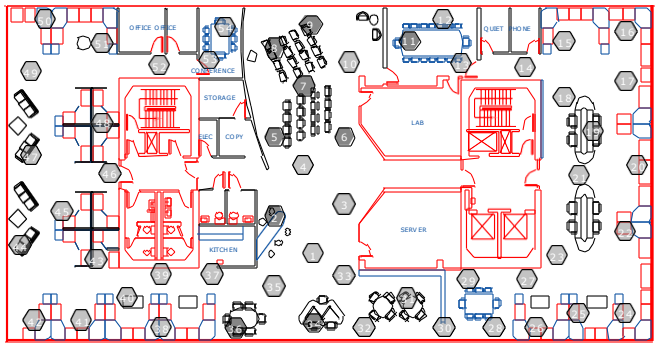


Temperature

72° F

Modeling sensor data

- Measure temperatures at some locations
- Predict temperatures throughout the environment

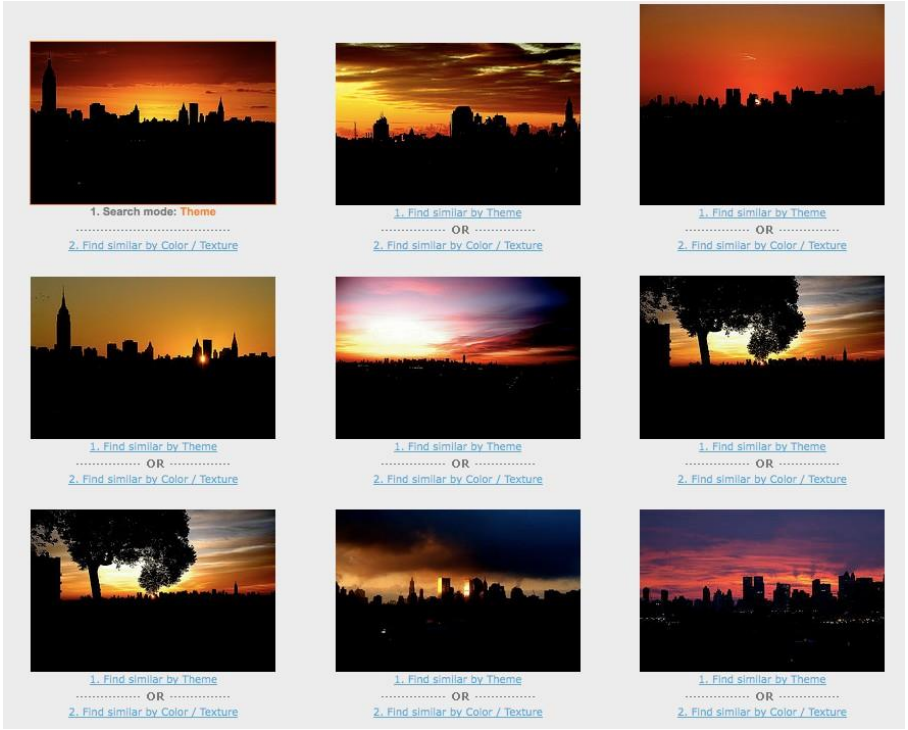


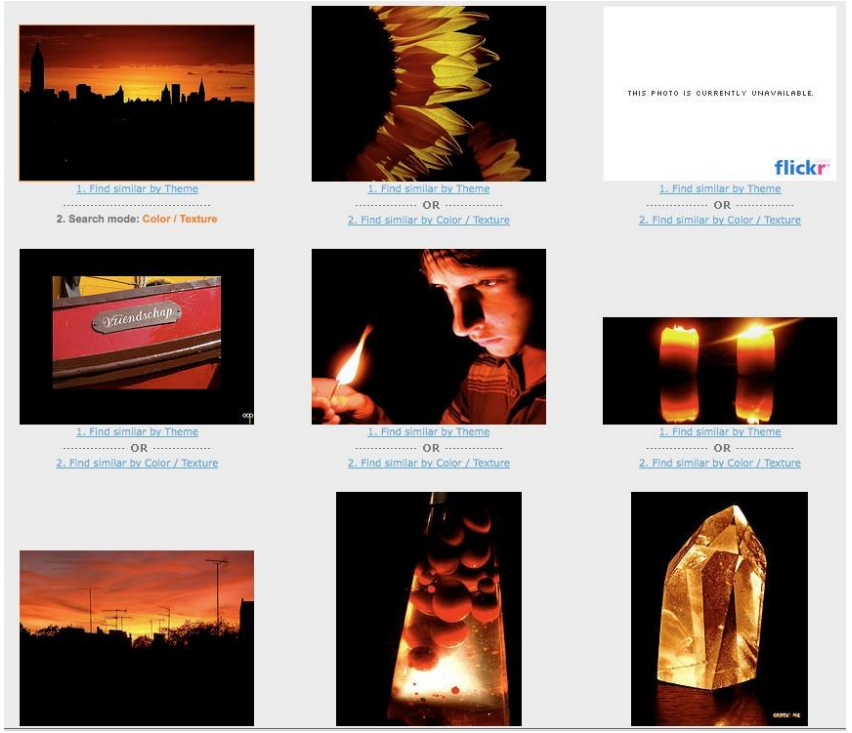
[Guestrin et al. ’04]

Similarity

finding data

Given image, find similar images





http://www.tiltomo.com/

Recommender System: Collaborative Filtering

See larger image

Share your own customer images

Publisher: learn how customers can search inside this book.

Please tell the publisher:

I'd like to read this book on Kindle

Don't have a Kindle? Get yours here.

Processing: A Programming Handbook for Visual Designers and Artists (Hardcover)

by Casey Reas (Author), Ben Fry (Author), John Maeda (Foreword)

(13 customer reviews)

Available from these sellers.

31 new from $47.95 **8 used** from $43.56

Get Free Two-Day Shipping

Get Free Two-Day Shipping for three months with a special extended free trial of Amazon Prime™. Add this eligible textbook to your cart to qualify. Sign up at checkout. See details.

Related Education & Training Services in Pittsburgh (What's this?) | Change location

Learn HTML Coding
www.FullSail.edu - Earn Your Bachelor's Degree in Web Design and Development.

Create Websites with HTML
http://www.unex.Berkeley.edu - Learn HTML Online, Start Anytime! with UC Berkeley Extension

Intensive XSLT Training
www.objectdatalabs.com/course10.asp - OnSite or in NYC, LA, SFO, ORD, DC Will customize & train as few as 3

Customers Who Bought This Item Also Bought

Processing: Creative Coding and Computational A... by Ira Greenberg
(7) $43.99

Visualizing Data: Exploring and Explaining Data... by Ben Fry
(11) $26.39

Making Things Talk: Practical Methods for Conne... by Tom Igoe
(15) $19.79

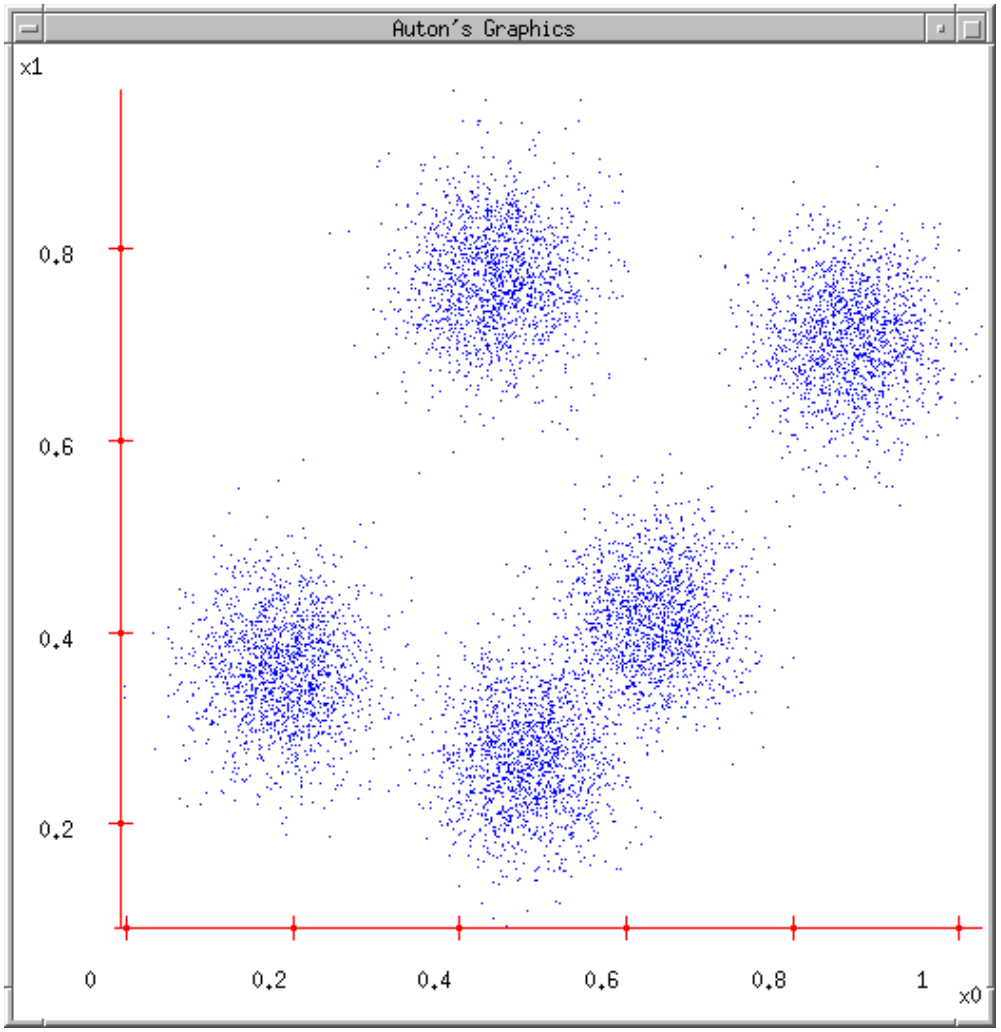
Physical Computing: Sensing and Controlling the... by Tom Igoe
(20) $19.00

Learning Processing: A Beginner's Guide to... by Daniel Shiffman
(7) $44.95

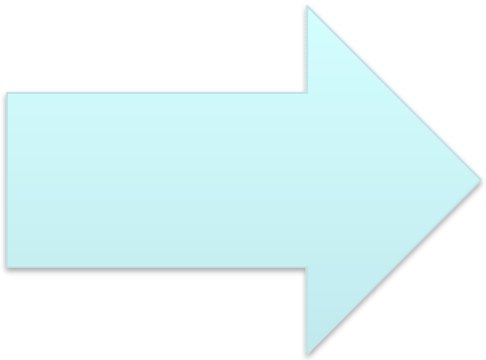
Clustering

discovering structure in data

Clustering Data: Group similar things

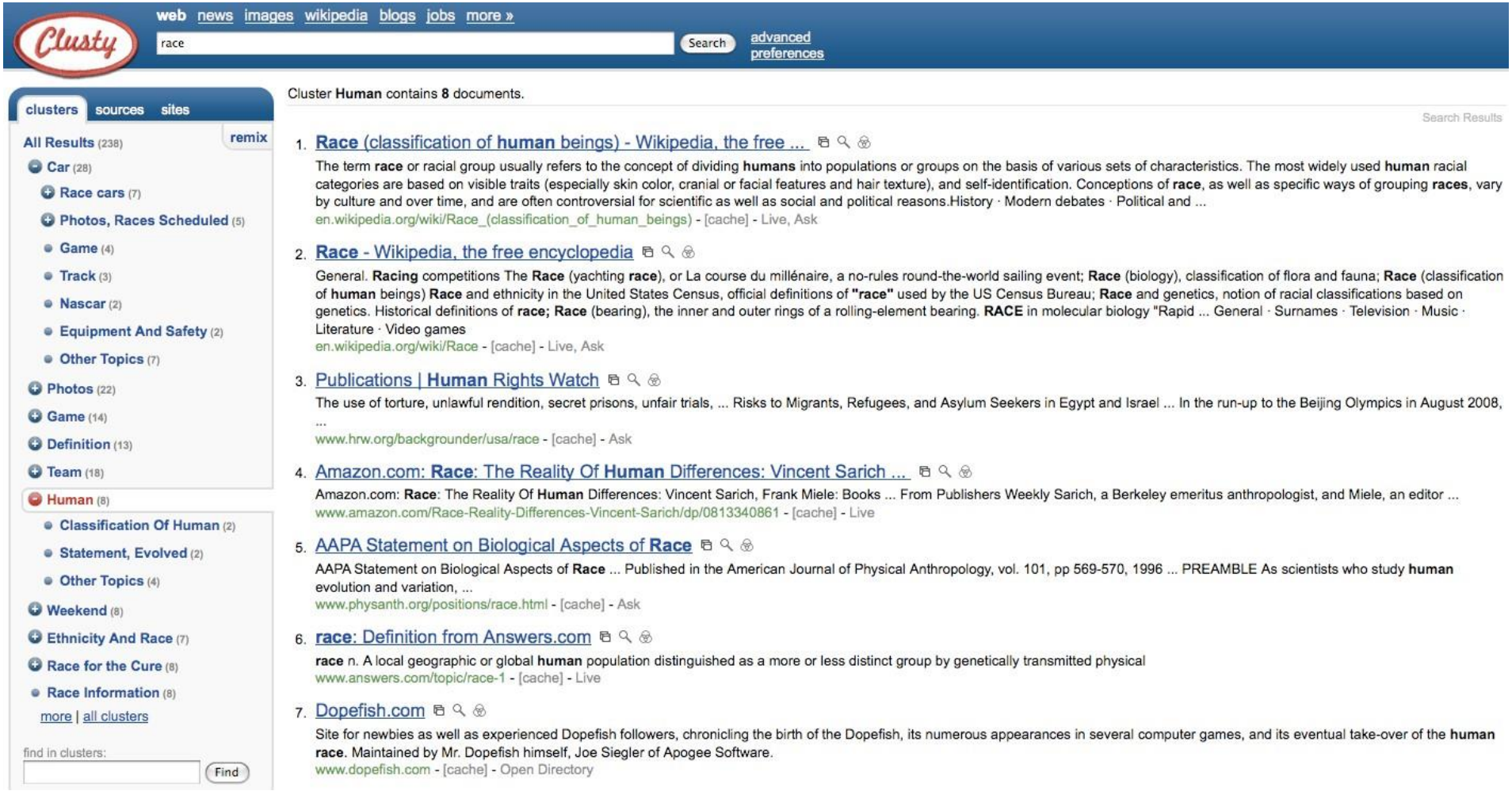


Clustering images



[Goldberger et al.]

Clustering web search results

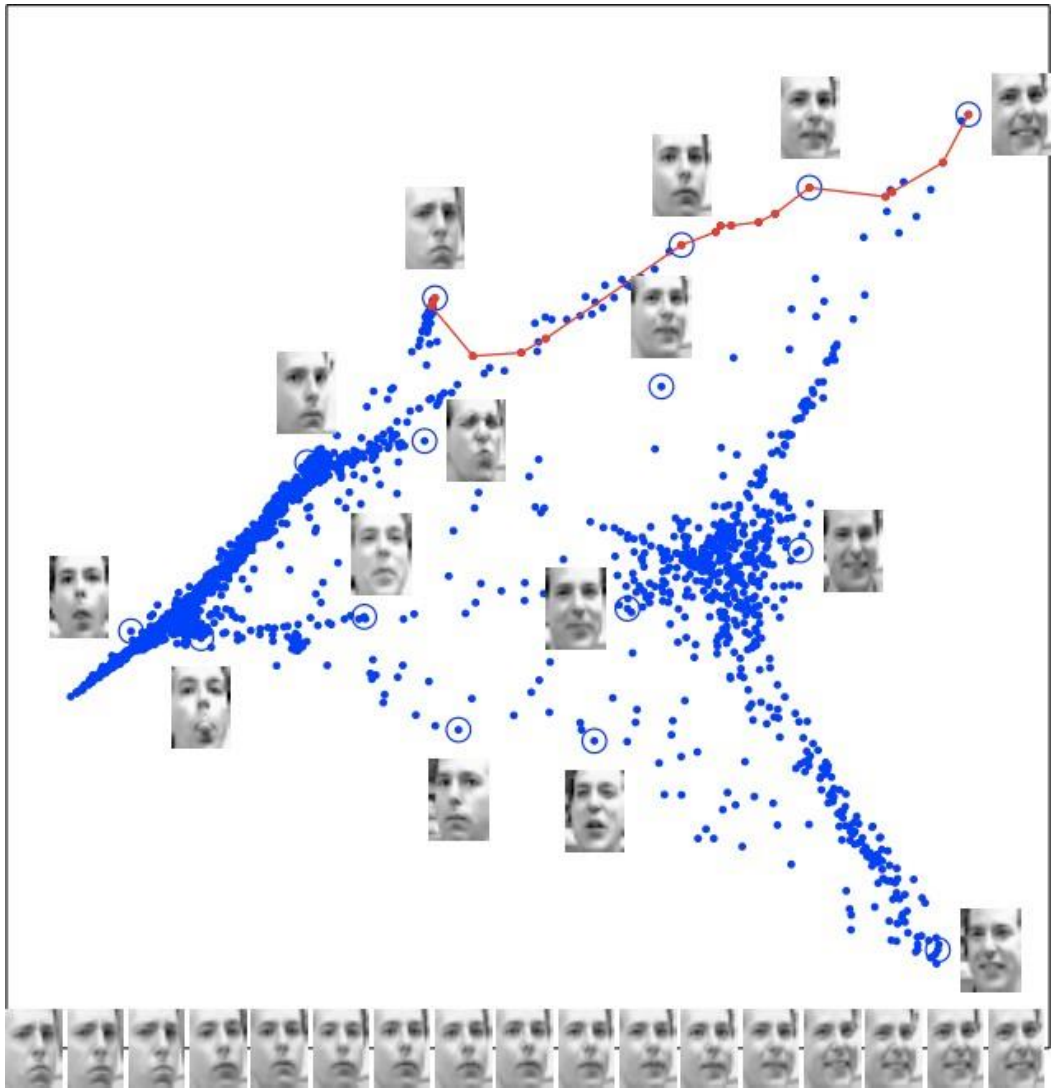


Embedding

discovering latent factor
visualizing data

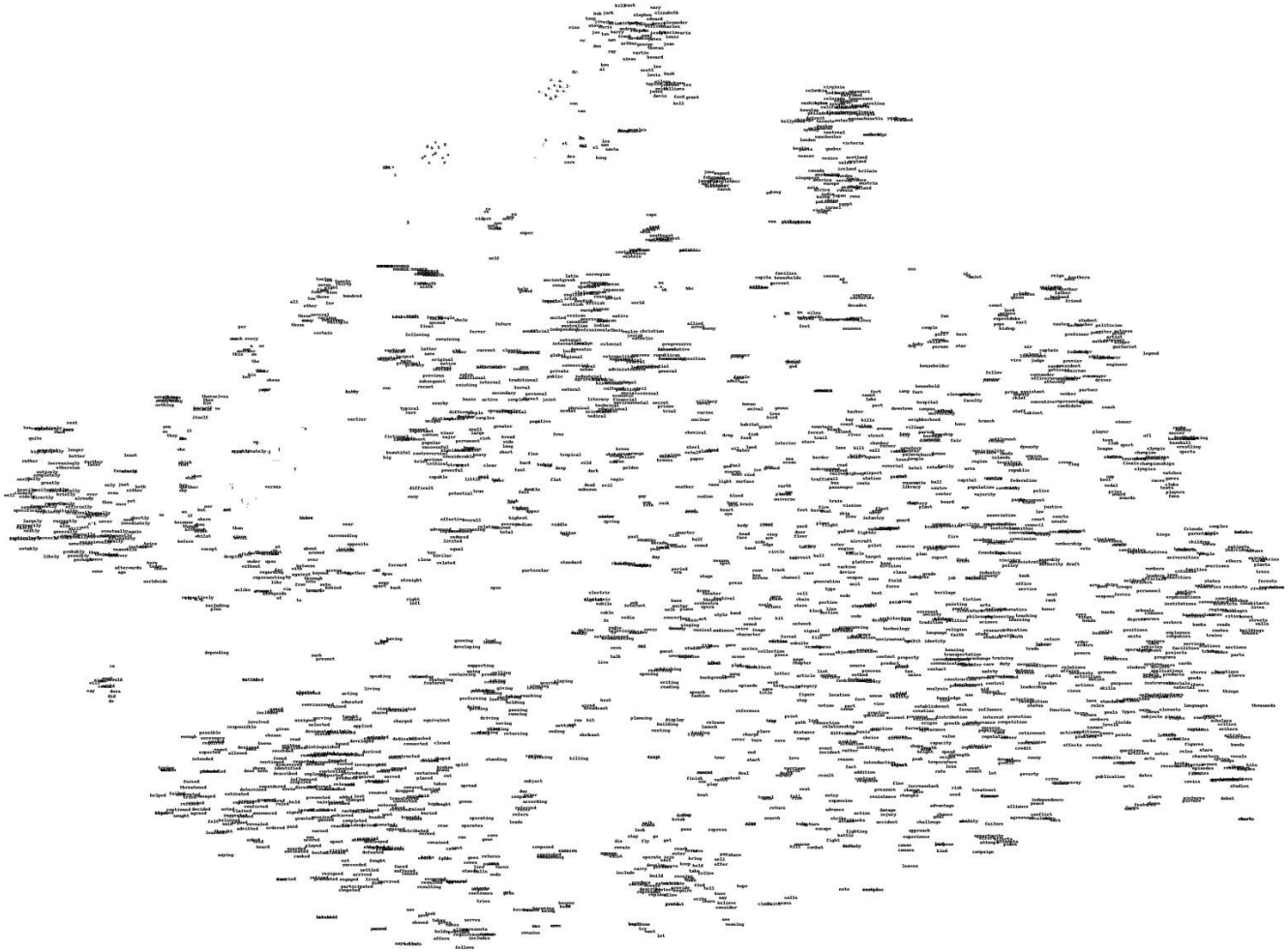
Embedding images

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



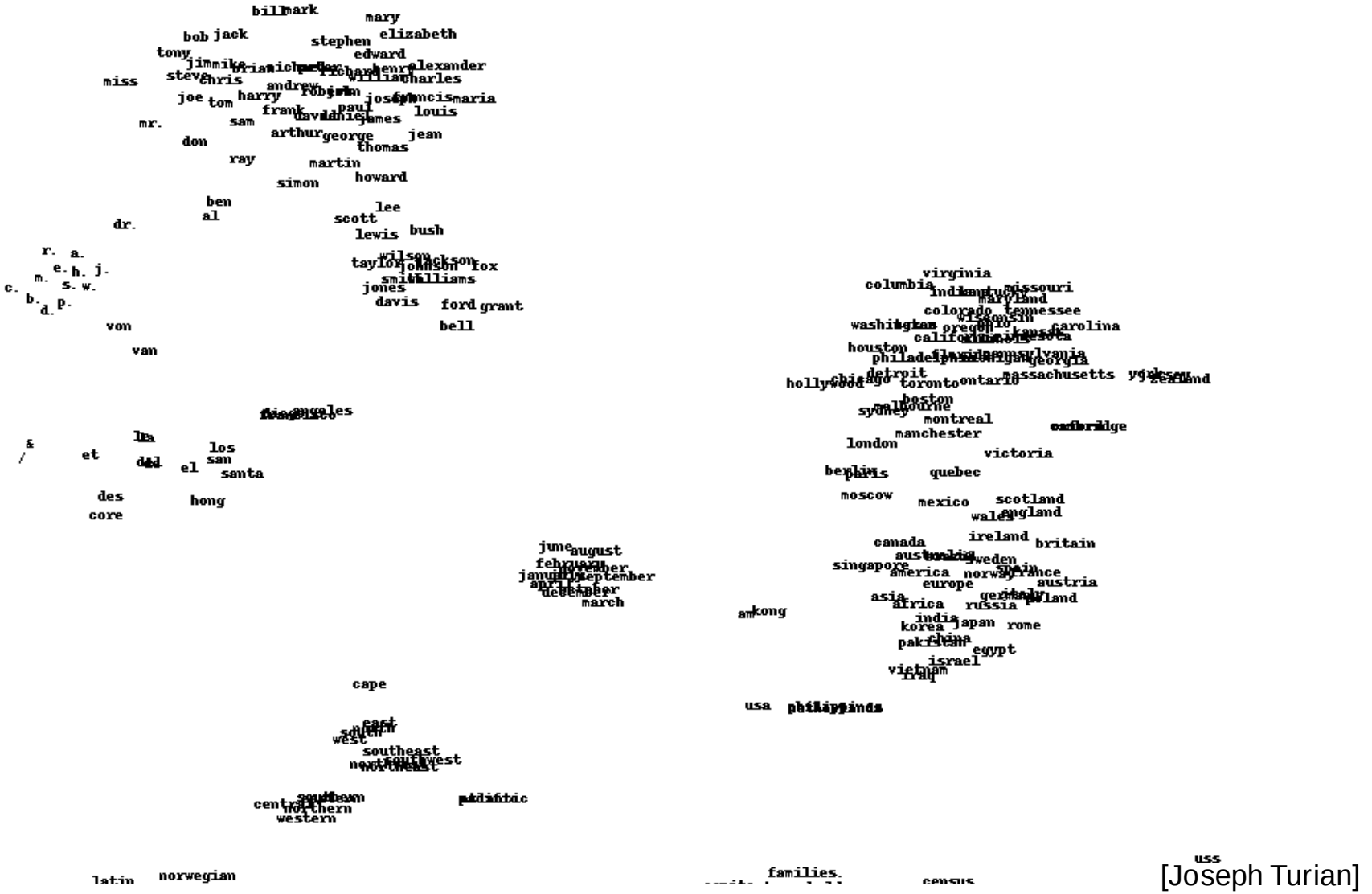
[Saul & Roweis ‘03]

Embedding words



[Joseph Turian]

Embedding words (zoom in)



[Joseph Turian]

Reinforcement Learning

training by feedback

Learning to act

- Reinforcement learning
- An agent
  - Makes sensor observations
  - Must select action
  - Receives rewards
    - positive for “good” states
    - negative for “bad” states

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

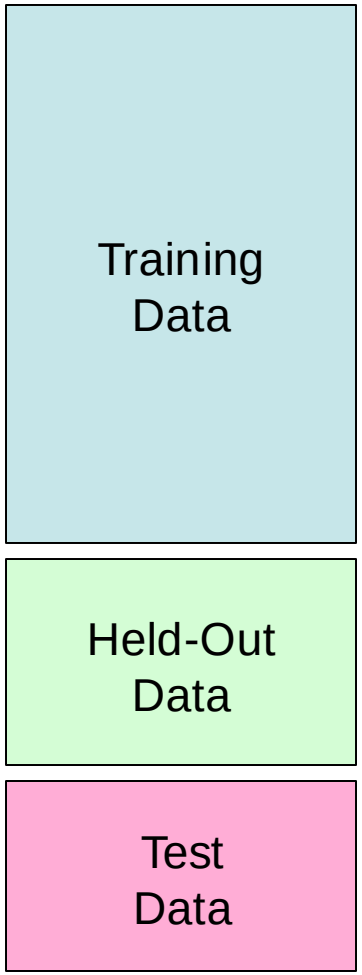

??

Machine Learning Categories

- Supervised Learning
  - Classification
  - Regression
- Unsupervised Learning
  - Clustering
  - Latent Factors
  - Matrix Completion
- Reinforcement Learning

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

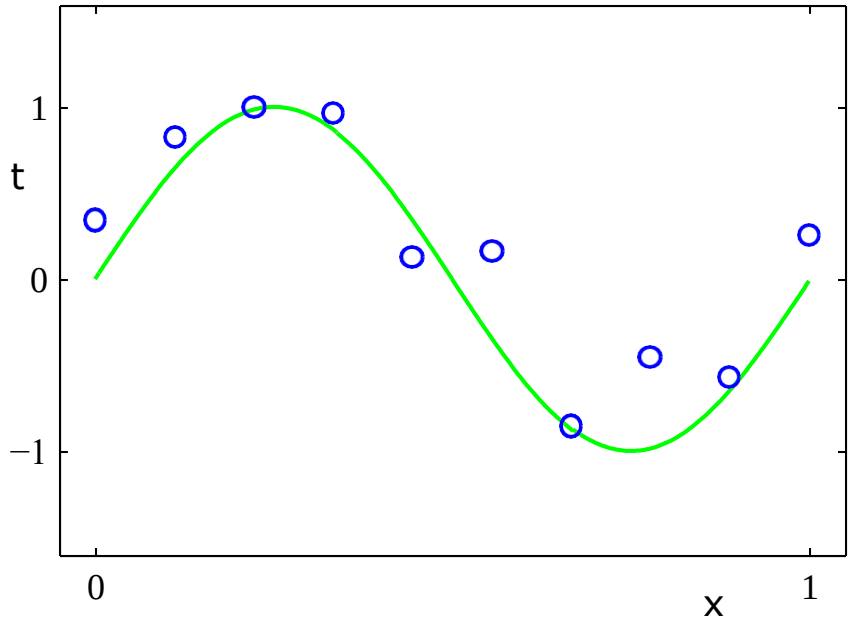


A Supervised Learning Example

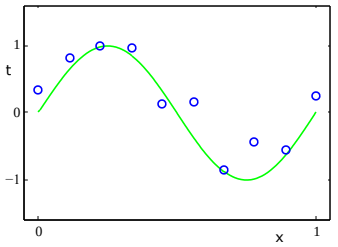
- Consider a simple, regression dataset:
  - $f : X \text{->} Y$
  - $X = \Re$
  - $Y = \Re$

- 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?

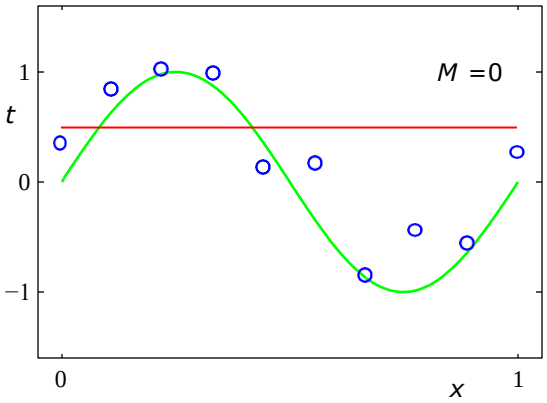
Dataset: 10 points generated from a sin function, with noise

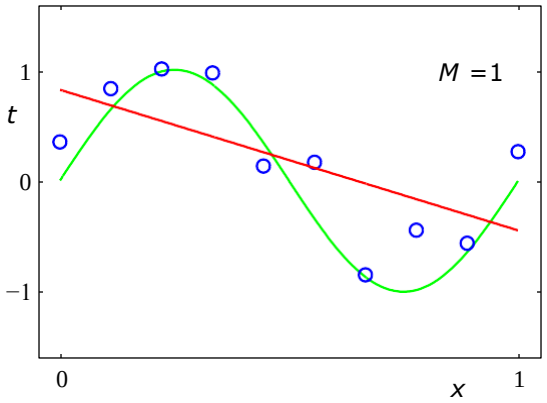


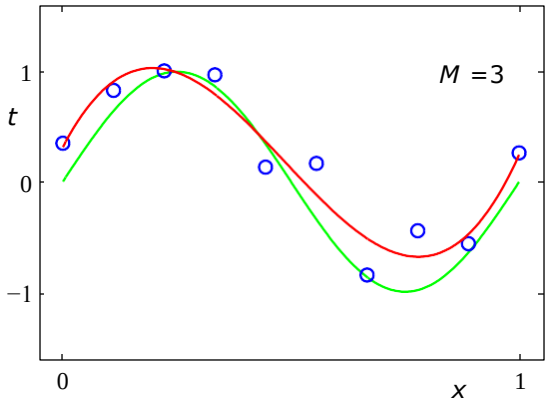
Hypo. Space: Degree-N Polynomials

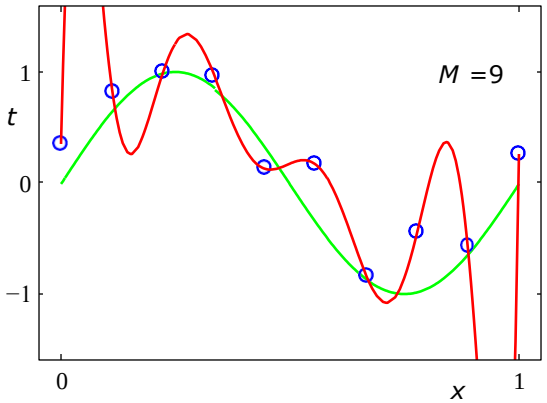


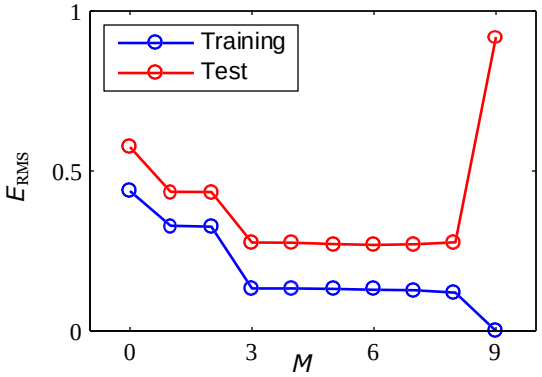
- Infinitely many hypotheses
- None / Infinitely many are consistent with our dataset
- How do we choose the best one?











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)

Checklist

1. Sign-up for Piazza
2. Sign-up for Tophat
3. Read the department's academic integrity policy
4. Reading materials: Murphy book Chapter 1