VIRGINIA TECH

Basic Understanding of Machine Learning

Lifu Huang

Computer Science, Virginia Tech

January 21, 2021

Slides adapted from Luke Zettlemoyer, David Sontag, Bert Huang

N L P L a b | Natural Language Processing Lab

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
- “Artificial intelligence will be the ultimate version of Google. The ultimate search engine that would understand everything on the web. It would understand exactly what you wanted, and it would give you the right thing. We’re nowhere near doing that now. However, we can get incrementally closer to that, and that is basically what we work on” – Larry Page, Co-founder and CEO, Google
- “Machine learning is going to result in a real revolution” – Greg Papadopoulos, CTO, Sun
- “Machine learning is today’s discontinuity” – Jerry Yang, CEO, Yahoo
- …

What is learning?

What is learning?










… … …

What is learning?










… … …

Which one is Cat??

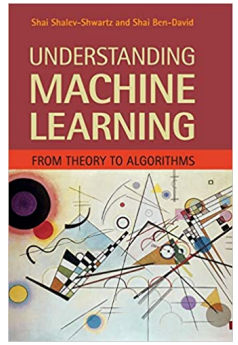
What's machine learning?

Machine learning (**ML**) is the study of computer algorithms that improve *automatically* through experience. It is seen as a subset of artificial intelligence. Machine learning algorithms build a model based on sample data, known as "training data", in order to make predictions or decisions *without being explicitly programmed* to do so.

WIKIPEDIA
The Free Encyclopedia

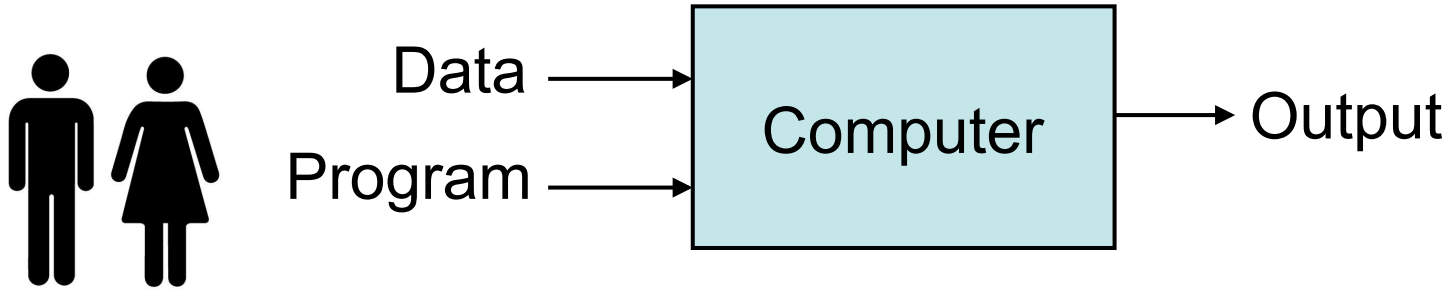
Machine learning refers to the *automated* detection of meaningful patterns in data.

Machine learning is the process of *converting experience into expertise or knowledge*. The input to a learning algorithm is training data, representing experience, and the output is some expertise

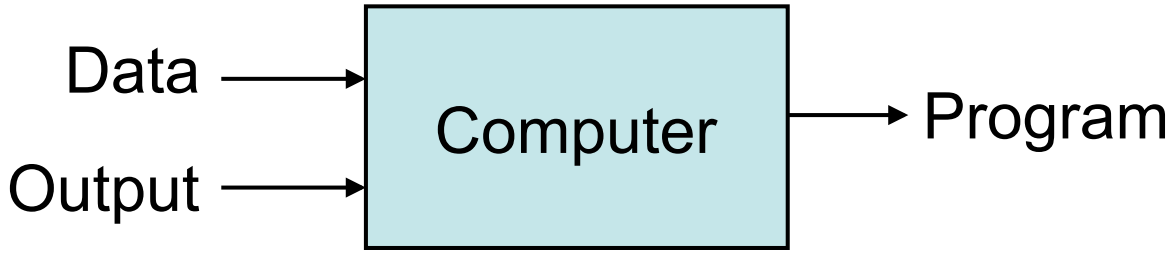


Traditional Programming v.s. Machine Learning

Traditional Programming



Machine Learning



Magic?

More like gardening

- **Seeds** = Algorithms
- **Nutrients** = Data
- **Gardener** = You
- **Plants** = Programs

When to apply machine learning?

- Tasks that are too complex to program
  - Tasks performed by animals/humans: *no well-defined program, e.g., driving, speech recognition, image understanding*
  - Tasks beyond human capabilities: *e.g., turning medical archives into medical knowledge, weather prediction, Web search engines*
- Adaptivity
  - Traditional programming: once been written down and installed, it stays unchanged
  - Machine learning: can adapt to the specific input data
  - Examples: decode handwritten text, speech recognition

When to apply machine learning?

- **A machine learning system can not predict stuff it does not know about**
- Let's say you teach an ML system about animals like this:

Number of Legs,	*Color,*	*Weight,*	*Animal*
4	Black	10KG	Dog
2	Orange	5KG	Chicken

If you now present it with a **Cow: 4 legs, black, and 200KG**, the machine learning system would predict it as "**Dog**". This is because it only knows about dogs and chickens and this was the closest match.

Types of machine learning problems

Supervised

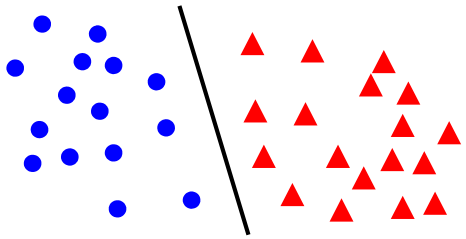
Learn through examples of which we know the desired output (what we want to predict), then generalize to the unseen test data.

Unsupervised

e.g., Is this a cat or a dog?

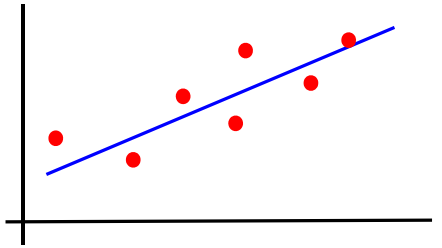
Are these emails spam or not?

Classification: from data to discrete classes



Reinforcement

Regression: predicting a numeric value



Types of machine learning problems

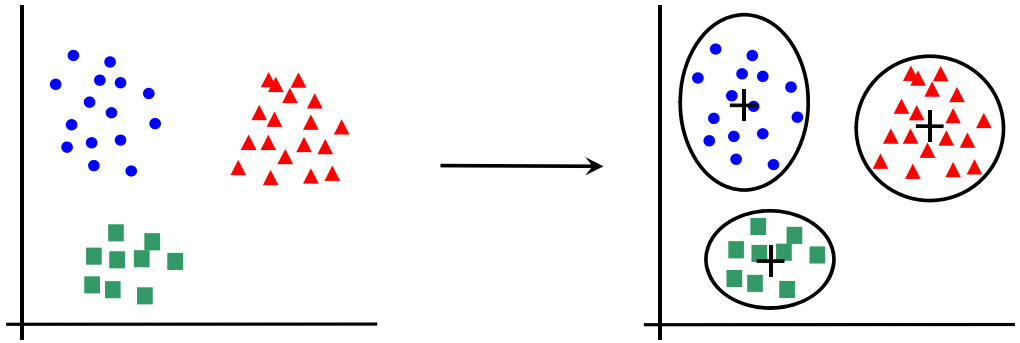
Supervised

This is no desired output. Learn something about the data, e.g., latent relationships, or compressed version of the data

e.g., I have photos and want to put them in 20 groups.

Unsupervised

Clustering: discovering structure in data



Reinforcement

Types of machine learning problems

Supervised

An agent (learner) interacts with an environment and watches the result of the interaction.

Unsupervised

Environment gives feedback via a positive or negative reward signal.

Reinforcement

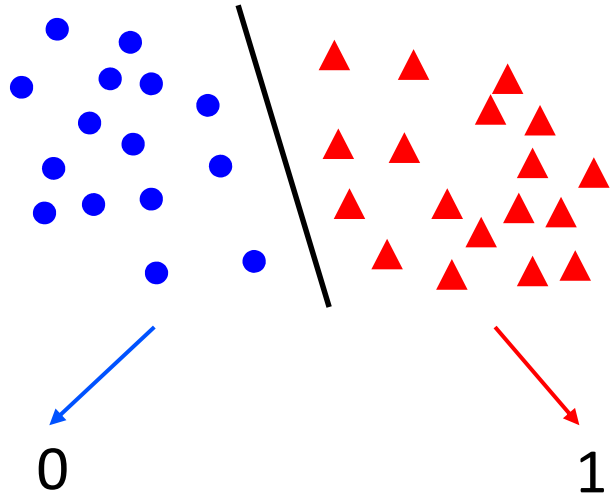
e.g., ***Playing chess****:* the only information available to the learner at training time is positions that occurred throughout actual chess games, labeled by who eventually won that game.

A good way of learning machine learning is to connect it with real-world problems that you are familiar with.

What is Machine Learning (by Examples)

Classification

from data to discrete classes



Spam filtering

data

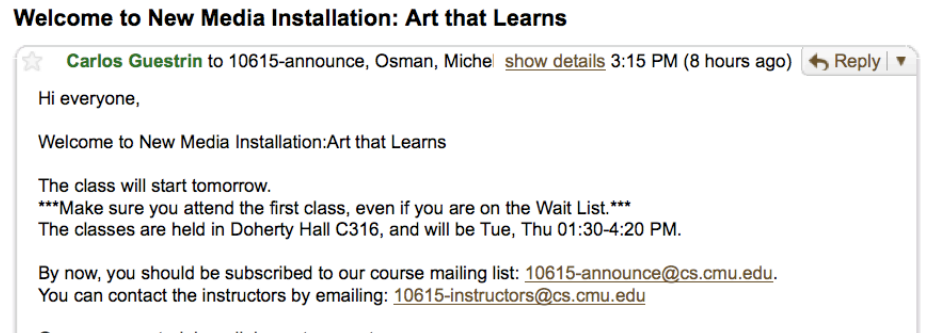
prediction

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

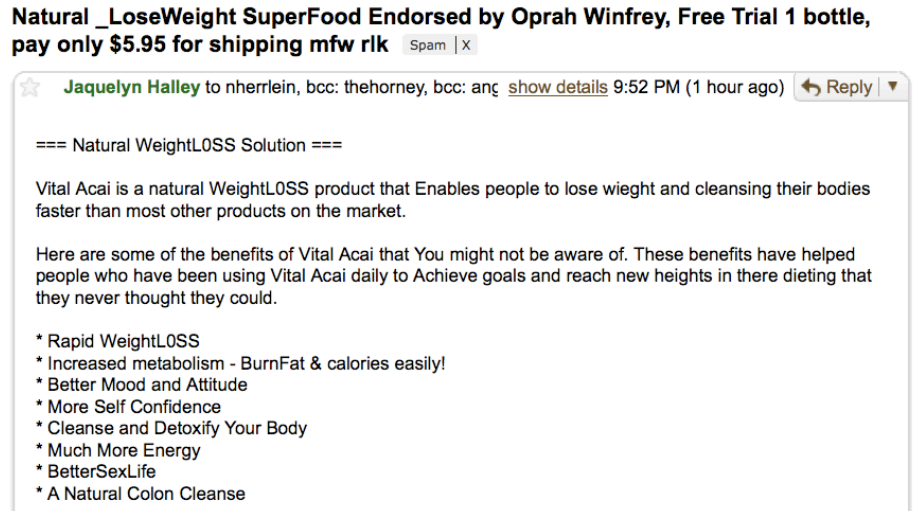
Carlos Guestrin to 10615-announce, Osman, Miche show details 3:15 PM (8 hours ago) Reply

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: ang 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.

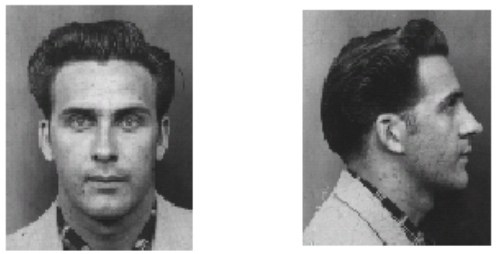
Here are some of the benefits of Vital Acai that You might not be aware of. These benefits have helped people who have been using Vital Acai daily to Achieve goals and reach new heights in there dieting that they never thought they could.

* 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

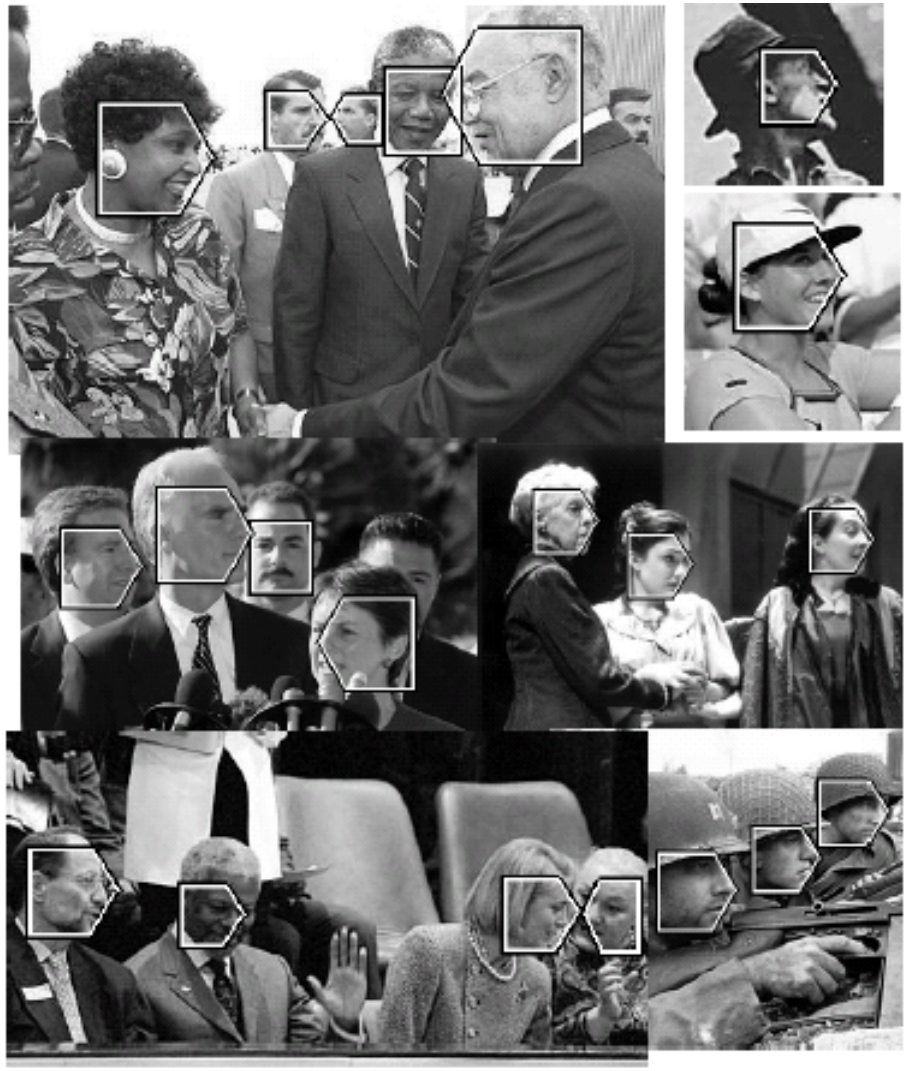
Spam
vs
Not Spam

Object detection

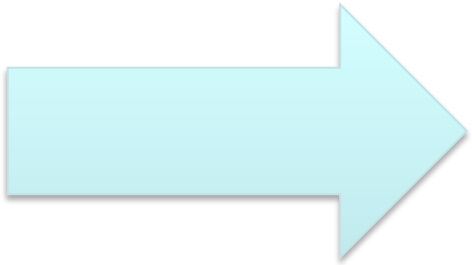
(Prof. H. Schneiderman)



Example training images
for each orientation



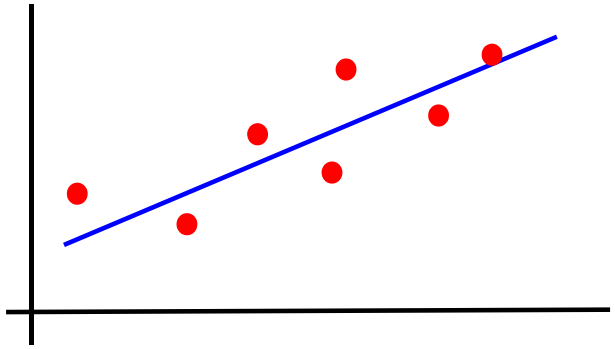
Weather prediction



What is Machine Learning (by Examples)

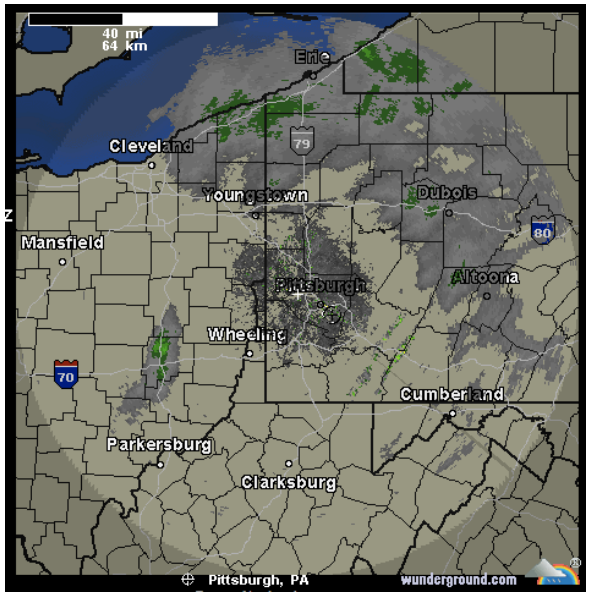
Regression

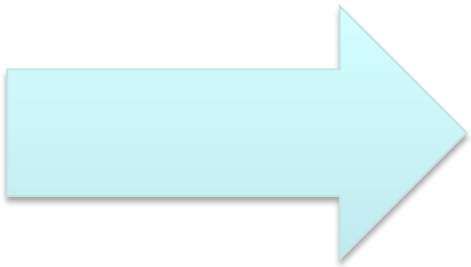
predicting a (continuous) numeric value



Stock market

Weather prediction revisted



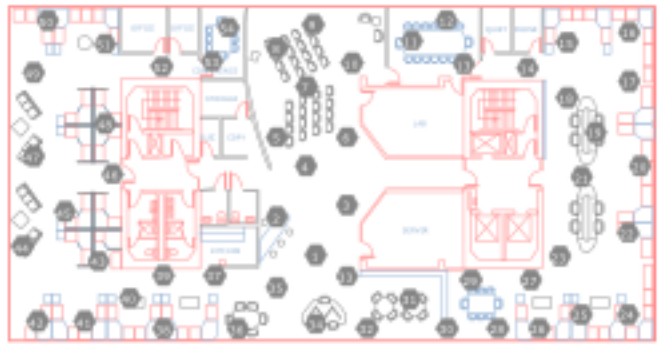


Temperature

72° F

Modeling sensor data

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



[Guestrin et al. '04]

What is Machine Learning (by Examples)

Similarity

finding relationship among data

Given an image, finding similar images

Input Image

1

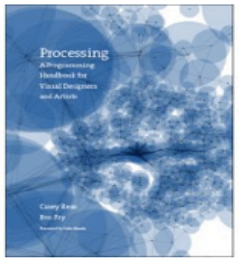
2

3

4

5

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

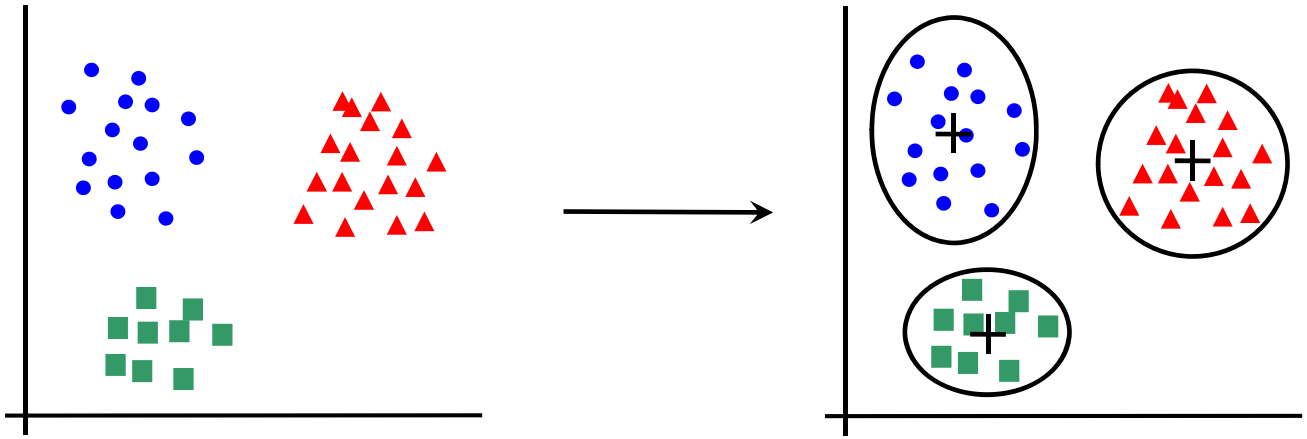
Physical Computing: Sensing and Controlling the... by Tom Igoe
(20) $19.00

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

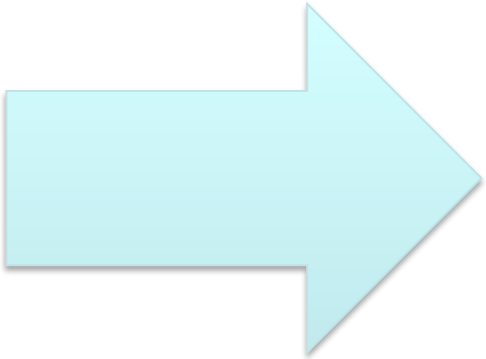
What is Machine Learning (by Examples)

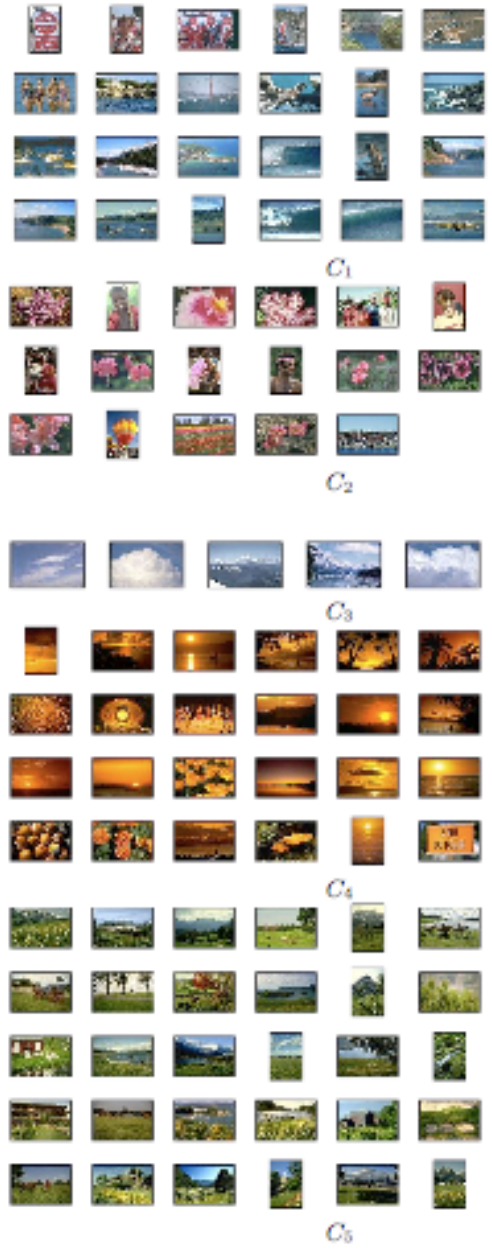
Clustering

discovering structure in data



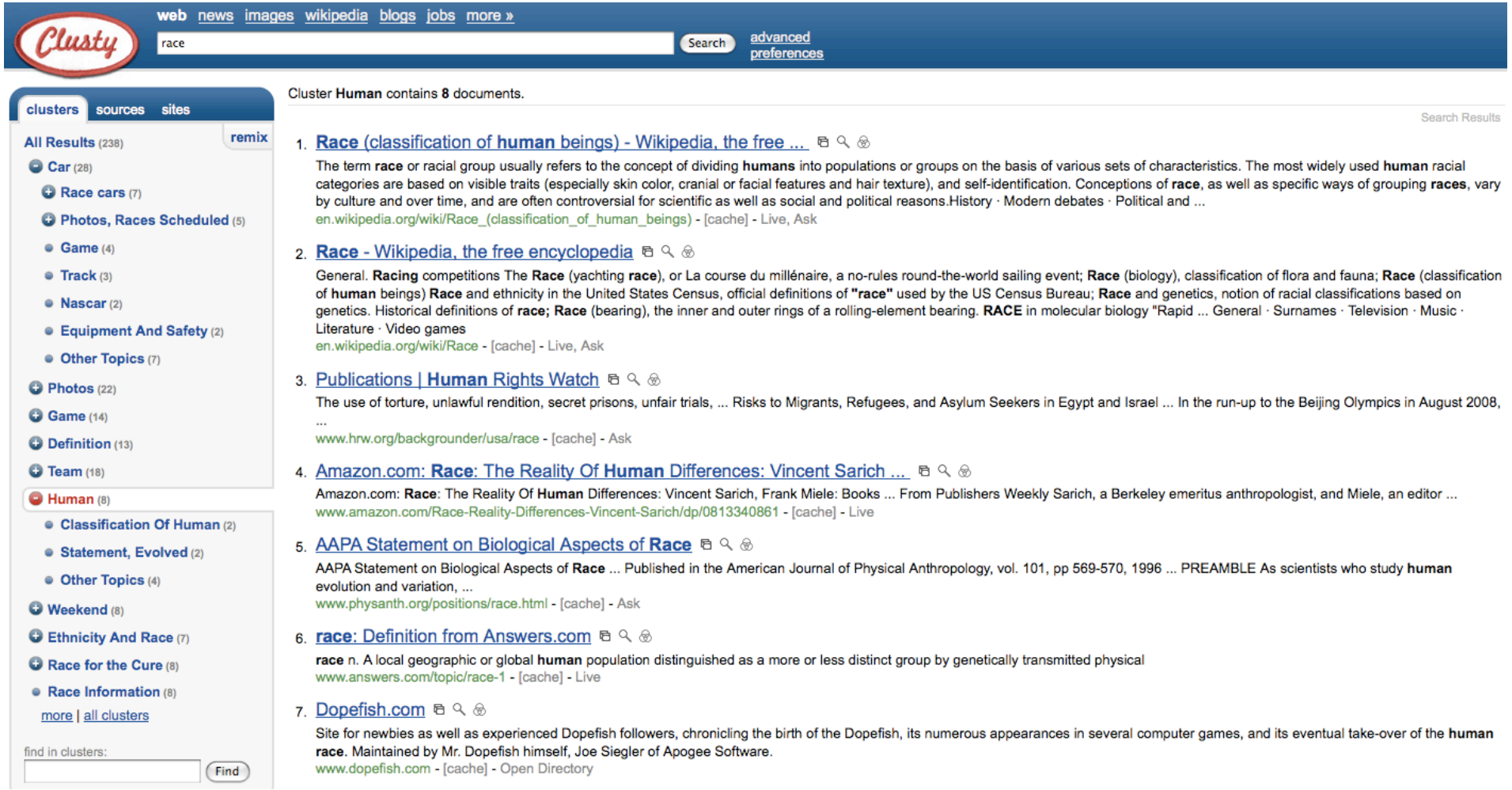
Clustering images





[Goldberger et al.]

Clustering web search results



What is Machine Learning (by Examples)

Reinforcement Learning

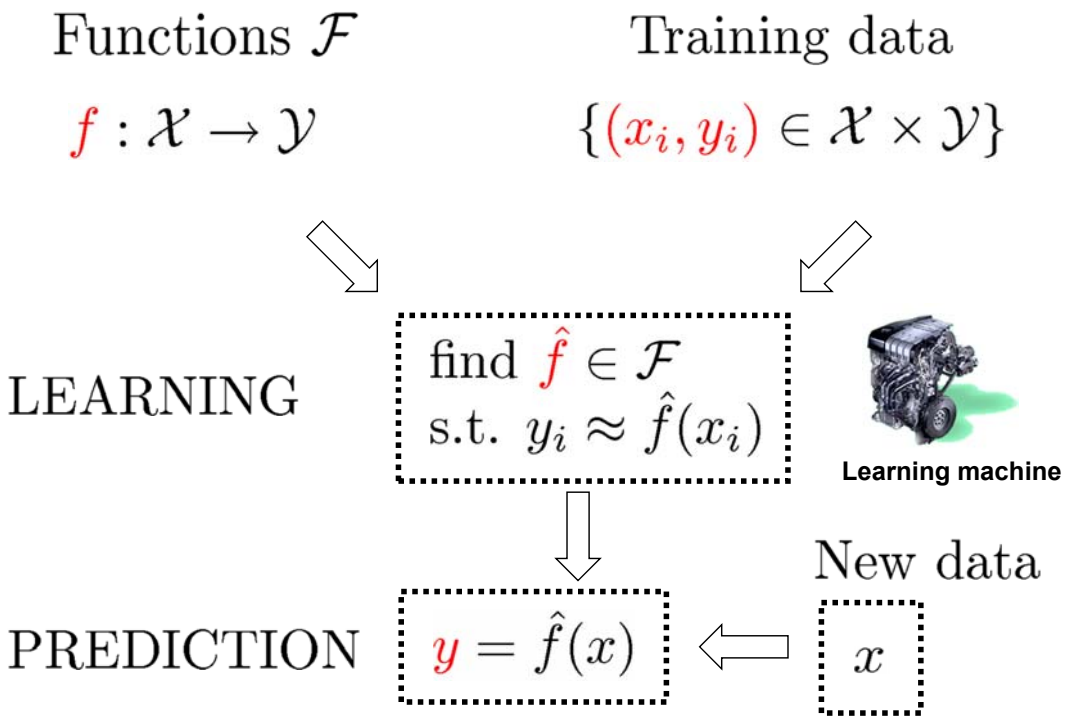
learning actions by feedback

Learning to act

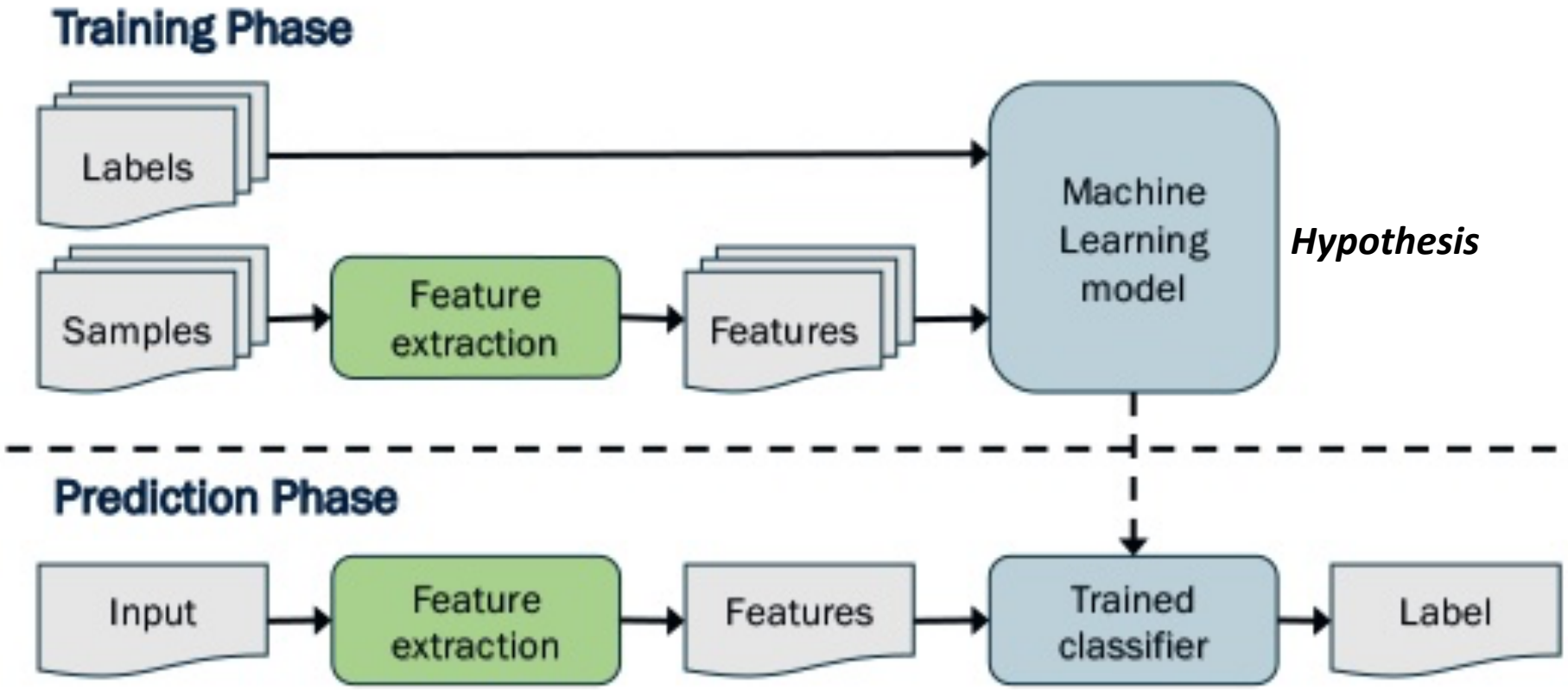
- Reinforcement learning
- An agent
  - Makes sensor observations
  - Must select action
  - Receives rewards
    - positive for “good” states
    - negative for “bad” states

Supervised Learning: find f

- Given: Training set $\{(x_i, y_i) | i = 1 \dots n\}$
- Learning: A good approximation to $f: X \to Y$
- 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 stocks to prices



Supervised Learning: find f



Example: Spam Filter

- Input: email
- Output: Spam/Not-spam
- Setup:
  - Get a large collection of example emails, each labeled as “spam” or “not-spam”
  - Note: someone has to manually label all the data to make sure the labels are (mostly) correct!!
  - Want to learn to predict labels of new, future emails
- Features: The attributes used to make the spam/not-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. …

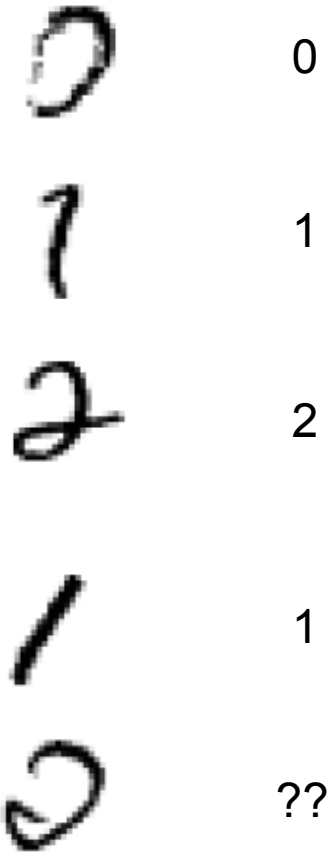
TO BE REMOVED FROM FUTURE MAILINGS, SIMPLY REPLY TO THIS MESSAGE AND PUT "REMOVE" IN THE SUBJECT.

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 manually label all the data to make sure the labels are (mostly) correct!!
  - Want to learn to predict labels of new, future digit images
- Features: The attributes used to make the digit decision
  - Pixels: Black/White
  - Shape Patterns: NumComponents, NumLoops
  - … …



Model Selection

- **Held-out Validation**
  - Sample an additional set of examples, independent of the training set, to validate the performance of the model

Model Selection

- **Held-out Validation**
  - Sample an additional set of examples, independent of the training set, to validate the performance of the model
- **K-Fold Cross Validation**
  - Randomly split the original training set into k folds
  - Each time, train model on k-1 folds and test the errors on the last one
  - Finally, the average of all test errors is the true error

k-Fold Cross Validation for Model Selection

input:
training set $S = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$
set of parameter values Θ
learning algorithm A
integer k
partition S into $S_1, S_2, \ldots, S_k$
foreach $\theta \in \Theta$
for $i = 1 \ldots k$
$h_{i,\theta} = A(S \setminus S_i; \theta)$
$\text{error}(\theta) = \frac{1}{k} \sum_{i=1}^{k} L_{S_i}(h_{i,\theta})$
output
$\theta^\star = \text{argmin}_\theta \left[\text{error}(\theta)\right]$
$h_{\theta^\star} = A(S; \theta^\star)$

Model Selection

- **Held-out Validation**
  - Sample an additional set of examples, independent of the training set, to validate the performance of the model
- **K-Fold Cross Validation**
  - Randomly split the original training set into k folds
  - Each time, train model on k-1 folds and test the errors on the last one
  - Finally, the average of all test errors is the true error
- **Best Practice** for Train-Validation-Test
  - Hold out test set completely hidden from training
  - Use validation on training data for model (or parameter) selection
  - Evaluate on held-out test data

***k*-Fold Cross Validation for Model Selection**

input:
training set $S = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$
set of parameter values Θ
learning algorithm A
integer k
partition S into $S_1, S_2, \ldots, S_k$
foreach $\theta \in \Theta$
for $i = 1 \ldots k$
$h_{i,\theta} = A(S \setminus S_i; \theta)$
$\text{error}(\theta) = \frac{1}{k} \sum_{i=1}^{k} L_{S_i}(h_{i,\theta})$
output
$\theta^\star = \text{argmin}_\theta \left[\text{error}(\theta)\right]$
$h_{\theta^\star} = A(S; \theta^\star)$

Basic Concepts

- **Data**: labeled instances, e.g., emails marked as spam or not-spam
  - Training Set
  - Held-out Set / Development Set / Validation Set
  - Test Set

Basic Concepts

- **Data**: labeled instances, e.g., emails marked as spam or not-spam
  - Training Set
  - Held-out Set / Development Set / Validation Set
  - Test Set
- **Features**: a list of values to characterize each instance while each value corresponds to a specific attribute
  - e.g., for spam filter task, the first value 0/1 indicates whether the email contains $.

Basic Concepts

- **Data**: labeled instances, e.g., emails marked as spam or not-spam
  - Training Set
  - Held-out Set / Development Set / Validation Set
  - Test Set
- **Features**: a list of values to characterize each instance while each value corresponds to a specific attribute
  - e.g., for spam filter task, the first value 0/1 indicates whether the email contains $.
- **Experiment Cycle**
  - Select a hypothesis f to best match training set
  - Tune hyperparameters on validation set
  - Compute accuracy of test set

Basic Concepts

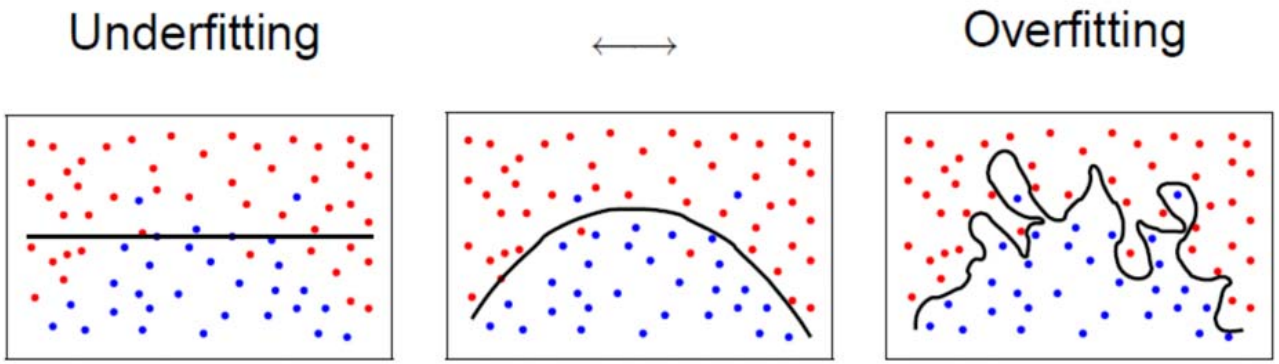
- **Data**: labeled instances, e.g., emails marked as spam or not-spam
  - Training Set
  - Held-out Set / Development Set / Validation Set
  - Test Set
- **Features**: a list of values to characterize each instance while each value corresponds to a specific attribute
  - e.g., for spam filter task, the first value 0/1 indicates whether the email contains $.
- **Experiment Cycle**
  - Select a hypothesis f to best match training set
  - Tune hyperparameters on validation set
  - Compute accuracy of test set
- **Evaluation**
  - Precision / Accuracy: fraction of instances predicted correctly

Basic Concepts

- **Generalization**
  - The real aim of supervised learning is to do well on test data which is unknown / unseen during learning
  - Choosing the values for the parameters that minimize the loss function on the training data is not necessarily the best policy
  - We want the learning machine to model the true regularities in the data and to ignore the noise in the data

Basic Concepts

- **Generalization**
  - The real aim of supervised learning is to do well on test data which is unknown / unseen during learning
  - Choosing the values for the parameters that minimize the loss function on the training data is not necessarily the best policy
  - We want the learning machine to model the true regularities in the data and to ignore the noise in the data
- **Overfitting**: fitting the training data very closely, but not generalizing well on test data
  - How to avoid overfitting: Add more data than the model "complexity" (Regularization)



Key Issues in Machine Learning

- What are good hypothesis spaces?
- How to find the best hypothesis?
- How to optimize for accuracy of unseen test data? (generalize well, avoid overfitting, etc.)
- Can we have confidence in results? How much data is needed?
- How to model applications as machine learning problems? (classification / regression / clustering / …)
- …