

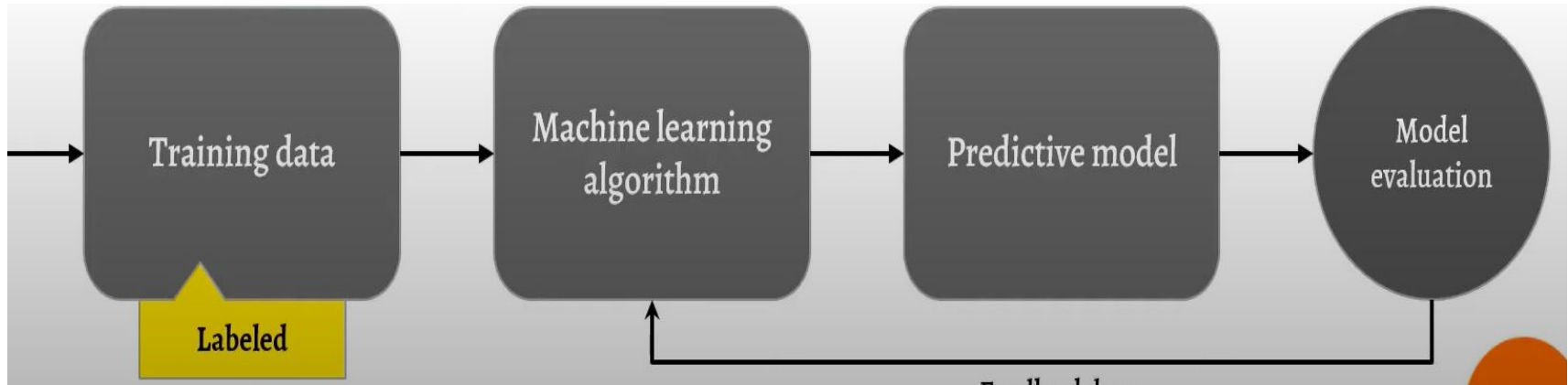
Semi-Supervised learning

Team members:

- Karan Pinakinbhai Jariwala
- Alihussain Ladiwala
- Varun Kumar Ejanthkar
- Kandi Sandeep Reddy

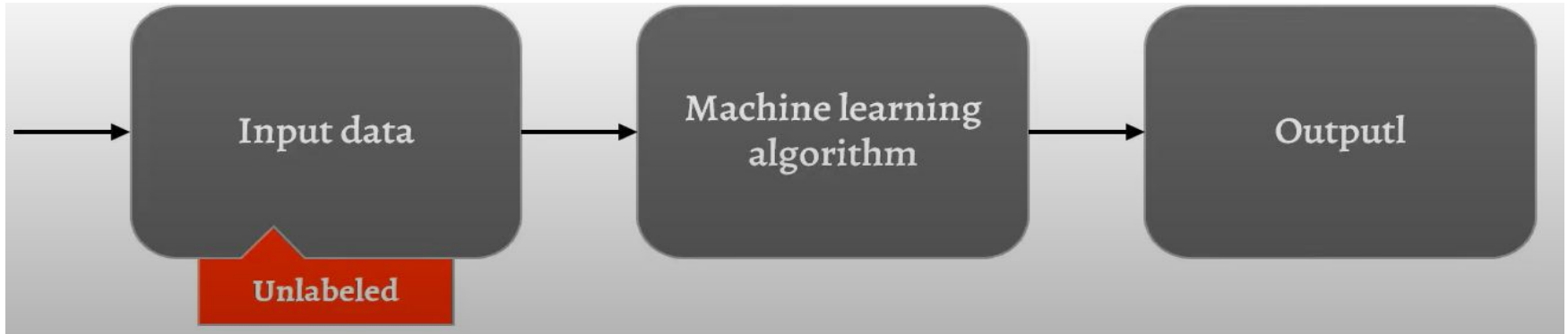
Introduction

- What is Supervised Learning?
 - Supervised Learning is a learning in which we teach or train the machine using data which are properly or rather labelled correctly.
 - Predicts a true value of y on the future data of x



Introduction

- What is Unsupervised Machine Learning
 - Unsupervised learning is the learning of machine using information that is neither classified nor labelled and allowing the algorithm to act on that information without guidance.
 - Divided into n instances and clustering is an example.



Drawbacks of these learning

- **Supervised Machine Learning** algorithm needs labeled data from which it will learn. We have enormous amount of data available in the world, including texts, images, audio, videos and many more, but only a tiny amount of data is actually labelled. This labeling of data is very costly process specially when dealing with large volume.
- **Unsupervised Machine Learning** algorithm has a disadvantage of application spectrum due to its unlabelled data.

Semi-Supervised Learning

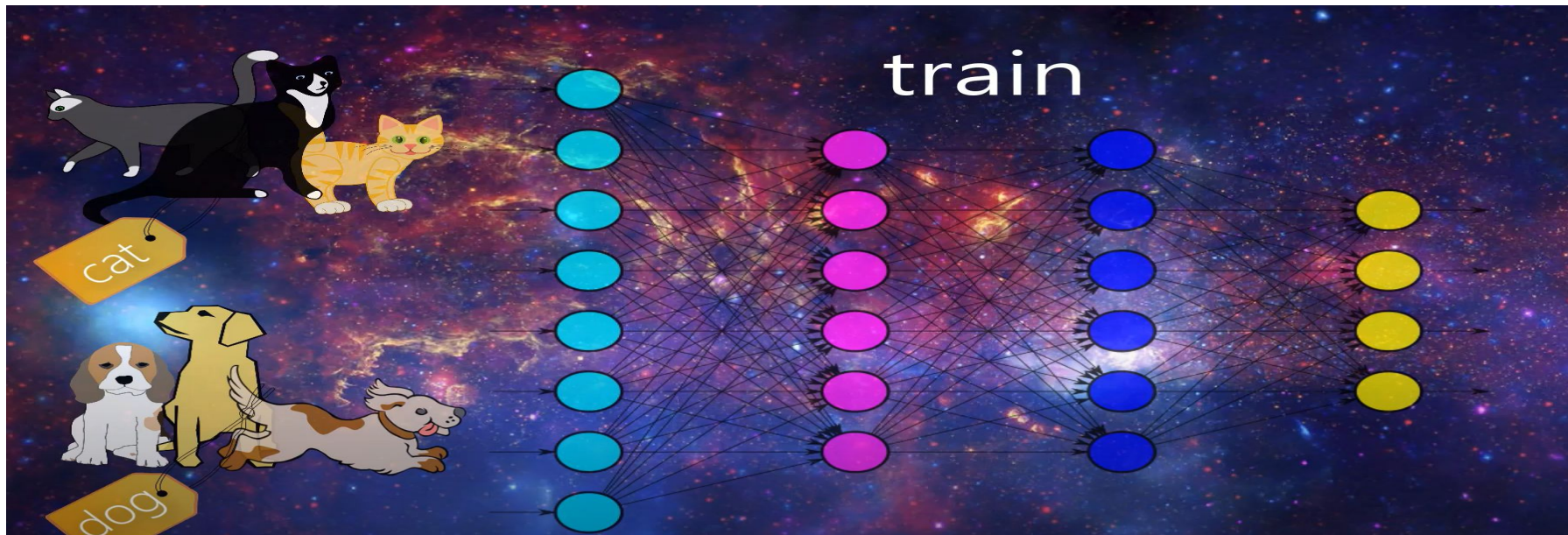
- Labeling a data can be very costly and manually impossible way to work with such large amounts of data.
- So the amalgamation of supervised and unsupervised learning is called semi-supervised machine learning.
- Semi-supervised machine learning is a combination of supervised and unsupervised learning. It uses a small amount of labeled data and a large amount of unlabeled data, which provides the benefits of both unsupervised and supervised learning while avoiding the challenges of finding a large amount of labeled data. That means you can train a model to label data without having to use as much labeled training data.

Semi-Supervised Learning



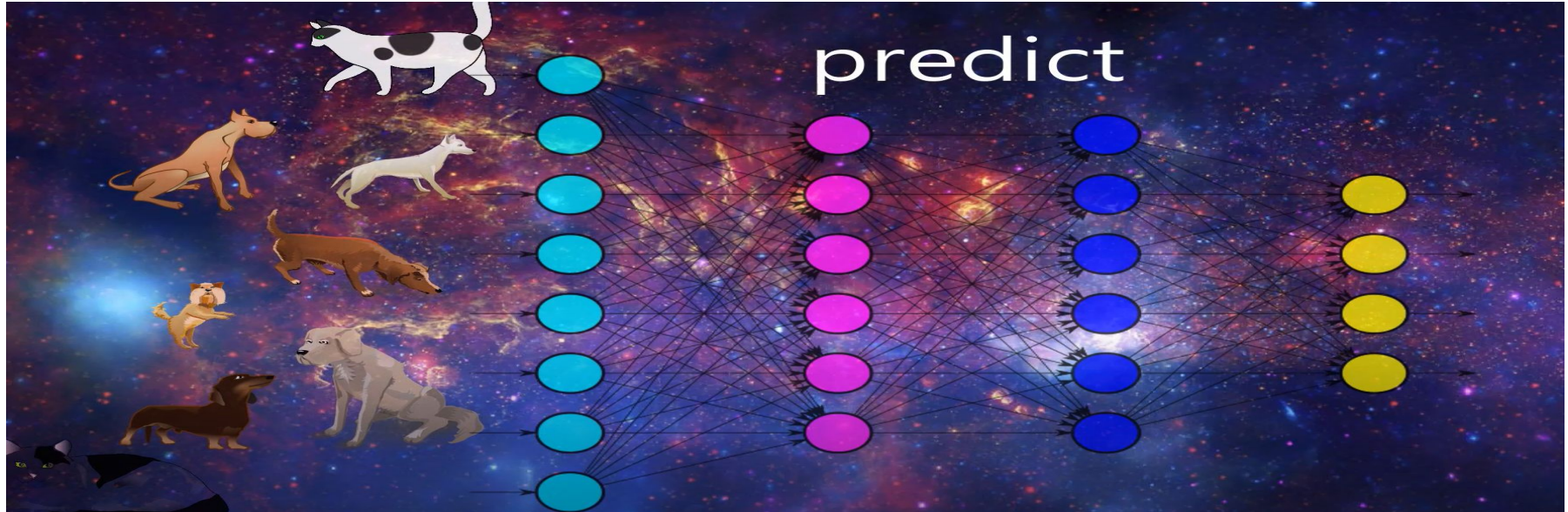
Pseudo-Labeling

We have some portion of dataset now this labeled data will be used as training set for our model basically what we do in regular supervised learning



Pseudo-Labeling

- After training we will use the model to predict our unlabeled data with individual output we predicted



Pseudo-Labeling

- Now we again train the model on full dataset which truly labeled and that was pseudo labeled



Steps for Semi Supervised Learning

- Train the model with the small amount of labeled training data just like you would in supervised learning, until it gives you good results.
- Then use it with the unlabeled training dataset to predict the outputs, which are pseudo labels since they may not be quite accurate.
- Link the labels from the labeled training data with the pseudo labels created in the previous step.
- Link the data inputs in the labeled training data with the inputs in the unlabeled data.
- Then, train the model the same way as you did with the labeled set in the beginning in order to decrease the error and improve the model's accuracy.

Applications

- **Internet Content Classification:**
 - Millions of web pages impossible to label
 - Semi-Supervised technique can help to classify the web pages
- **Audio/Video analysis:**
 - Overwhelming amount of audio and videos are present.
 - Not feasible to label the massive amount of data
 - Semi-Supervised technique can help to classify the Audio/Video analysis
- **Protein Sequence Classification:**
 - DNA strands are typically very large in size, the rise of Semi-Supervised Learning has been imminent in this field

Co-training


What is Co-training?

Many problems have two different sources of info you can use to determine label.

Eg: classifying web pages: can use words on page or words on links pointing to the page

Prof. Avrim Blum My Advisor

Avrim Blum's home page Page 1 of 1



Avrim Blum
Professor of Computer Science
Department of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213-5881
avrim@cs.cmu.edu

Office: Wean 4130
Tel: (412) 268-4042
Fax: (412) 268-5576
Address: Nuclei Stenger, Wean 4136, 268-2779

Check out our new faculty members Ryan O'Donnell and Luis von Ahn.

My main research interests are machine learning theory, approximation algorithms, on-line algorithms, and algorithmic game theory. I was on the Program Committee for FOCS 2009 (Simp. Foundations of Computer Science), ACM-ECC 2008 (Electronic Commerce), and FSTCS 2007 (Foundations of Theoretical Computer Science), and was recently local organizer for COLT 2008 and FOCS 2007. I also co-organized the 2007 Foundations of Computational Mathematics Workshop on Algorithms, Game Theory and Metric Embeddings. A while back I served as Program Chair for FOCS 2005 and I've done some work in AI Planning. For more information on my research, see the publications and research interests links below. I am also affiliated with the Machine Learning department.

I am currently (Spring 2009) teaching 15-079(S): Machine Learning Theory.


- Publications
- Research Interests
- Survey Talks
- Thesis Seminars, Theory Lunch ML lunch
- Conferences
- Family pictures, Other pictures, My Startup Page
- My Tutorial on Machine Learning Theory given at FOCS 2005 and a short story.

My advisors: Avrim Blum, Ramona Ognjen, Vlado Radoje, Miroslav Radoje, Miroslav Radoje, Miroslav Radoje

x - Link info & Text info

Prof. Avrim Blum My Advisor

Avrim Blum's home page Page 1 of 1



x₁ - Link info

Prof. Avrim Blum My Advisor

Avrim Blum's home page Page 1 of 1



Avrim Blum
Professor of Computer Science
Department of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213-5881
avrim@cs.cmu.edu

Office: Wean 4130
Tel: (412) 268-4042
Fax: (412) 268-5576
Address: Nuclei Stenger, Wean 4136, 268-2779

Check out our new faculty members Ryan O'Donnell and Luis von Ahn.

My main research interests are machine learning theory, approximation algorithms, on-line algorithms, and algorithmic game theory. I was on the Program Committee for FOCS 2009 (Simp. Foundations of Computer Science), ACM-ECC 2008 (Electronic Commerce), and FSTCS 2007 (Foundations of Theoretical Computer Science), and was recently local organizer for COLT 2008 and FOCS 2007. I also co-organized the 2007 Foundations of Computational Mathematics Workshop on Algorithms, Game Theory and Metric Embeddings. A while back I served as Program Chair for FOCS 2005 and I've done some work in AI Planning. For more information on my research, see the publications and research interests links below. I am also affiliated with the Machine Learning department.

I am currently (Spring 2009) teaching 15-079(S): Machine Learning Theory.

- Publications
- Research Interests
- Survey Talks
- Thesis Seminars, Theory Lunch ML lunch
- Conferences
- Family pictures, Other pictures, My Startup Page
- My Tutorial on Machine Learning Theory given at FOCS 2005 and a short story.

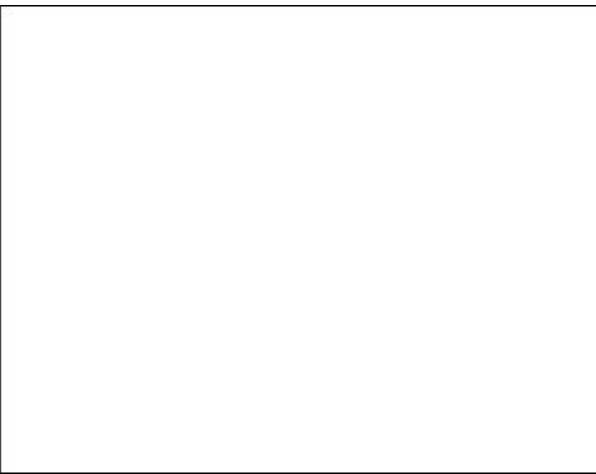
My advisors: Avrim Blum, Ramona Ognjen, Vlado Radoje, Miroslav Radoje, Miroslav Radoje, Miroslav Radoje

x₂ - Text info

Co-training

Idea: Use small labeled sample to learn initial rules.

- E.g., “my advisor” pointing to a page is a good indicator it is a faculty home page.
- E.g., “I am teaching” on a page is a good indicator it is a faculty home page.



my advisor



Avrim Blum's home page Page 1 of 1



Avrim Blum
Professor of Computer Science
Department of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213-3891
avrim@cs.cmu.edu
Office: Wean 4130
Tel: (412) 268-6452
Fax: (412) 268-8576
Admin assist: Nicole Stenger, Wean 4116, 268-3779

Check out our new faculty members [Ryan O'Donnell](#) and [Luis von Ahn](#).

My main research interests are machine learning theory, approximation algorithms, on-line algorithms, and algorithmic game theory. I was/am on the Program Committees for FOCS 2008 (Symp. Foundations of Computer Science), ACM-EC 2008 (Electronic Commerce), and COLT 2007 (Conference on Learning Theory), and was recently local organizer for COLT 2006 and FOCS 2005. I also co-organized the 2005 Foundations of Computational Mathematics Workshop on Algorithmic Game Theory and Metric Embeddings. A while back I served as Program Chair for FOCS 2000 and I've done some work in AI Planning. For more information on my research, see the publications and research interests links below. I am also affiliated with the Machine Learning department.

I am currently (Spring 2008) teaching 15-859/901 Machine Learning theory

Publications

Research Interests

Survey Talks

Courses

ALADDIN: Algorithms and Complexity Group

ACO Program Home Page

Theory Seminars, Theory lunch ML lunch

Family pictures, Other pictures, My Startup Page

My Tutorial on Machine Learning Theory given at FOCS 2003 and a short essay.

My advisees: Aaron Roth, Katrina Ligett, Nina Balcan, Magzi Robert Rowhangira, Shobha Venkataraman.

Past advisees: Prasad Chalasani, Santosh Vempala, Carl Burch, Adam Kalai, John Langford, Nikhil Bansal, Martin Zinkevich, Shuchi Chawla, Brendan McMahan.

Previous

Co-training

Idea: Use small labeled sample to learn initial rules.

- E.g., “my advisor” pointing to a page is a good indicator it is a faculty home page.
- E.g., “I am teaching” on a page is a good indicator it is a faculty home page.

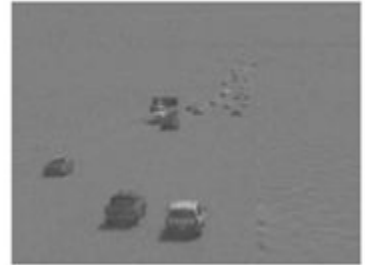
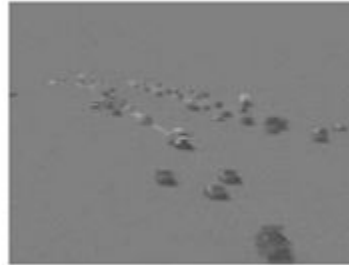
Then look for unlabeled examples where one rule is confident and the other is not. Have it label the example for the other.

Training 2 classifiers, one on each type of info. Using each to help train the other.

Co-training

Another Example:

E.g. identifying objects in images. Two different kinds of preprocessing.



Co-training

- Setting is each example $x = \langle x_1, x_2 \rangle$, where x_1, x_2 are two “views” of the data.
- Have separate algorithms running on each view. Use each to help train the other.
- Basic hope is that two views are consistent. Using agreement as proxy for labeled data.

Co-training (metrics)

- Images with 50 labeled cars. 22,000 unlabeled images.
- Factor 2-3+ improvement.

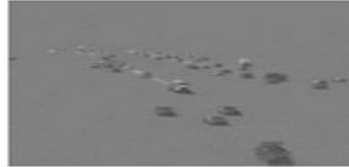
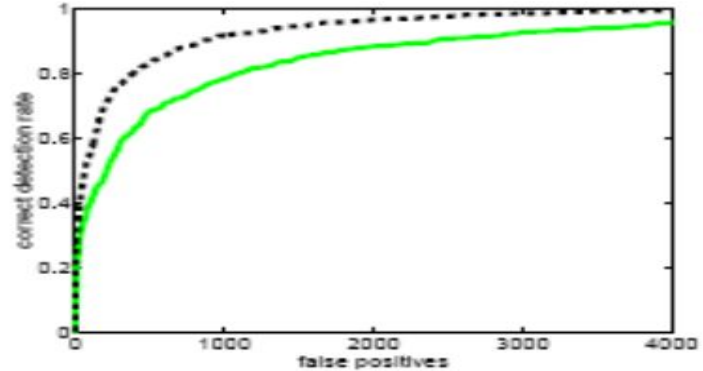


Figure 1: Example images used to test and train the car detection system. On the left are the original images. On the right are background subtracted images.

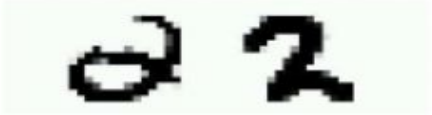

Graph-based methods

Graph-based methods

- Suppose we believe that very similar examples probably have the same label.
- If you have a lot of labeled data, this suggests a Nearest-Neighbor type of alg.
- If you have a lot of unlabeled data, perhaps can use them as “stepping stones”

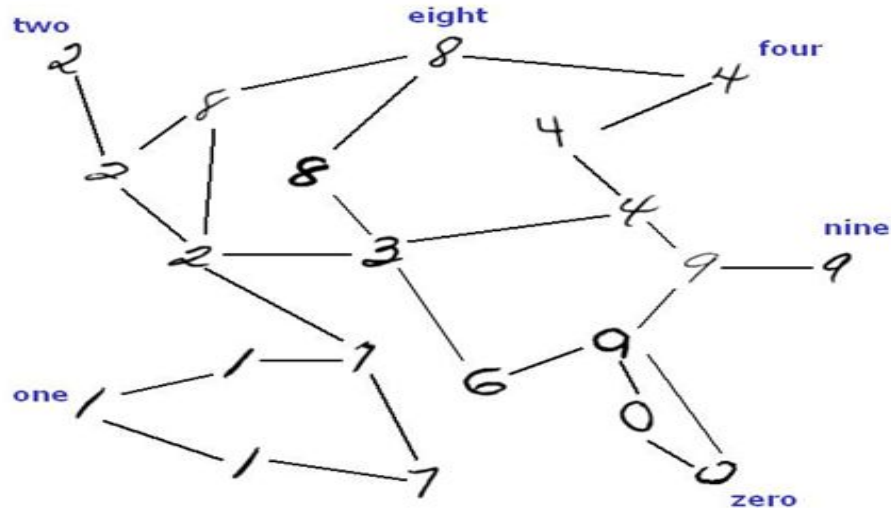
Graph-based methods

Popular Optical Character Recognition tools such as Tesseract, Kofax uses Graph based semi-supervised learning to extract the handwritten text.

 <p>not similar</p>	 <p>'indirectly' similar with stepping stones</p>
--	---

Graph-based methods

- Idea: construct a graph with edges between very similar examples.
- Unlabeled data can help “glue” the objects of the same class together.



Graph-based methods

- Idea: construct a graph with edges between very similar examples.
- Unlabeled data can help “glue” the objects of the same class together.



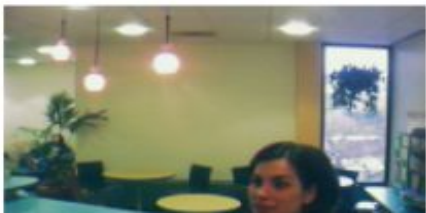
image 4005



neighbor 1: time edge



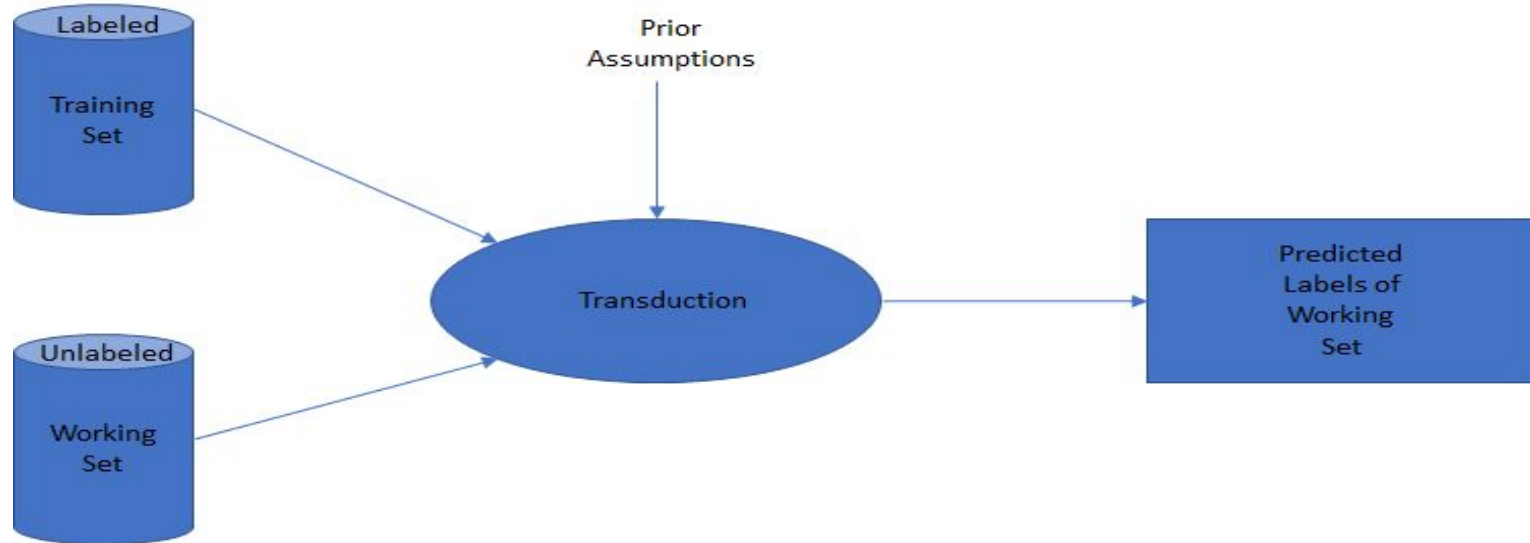
neighbor 2: color edge



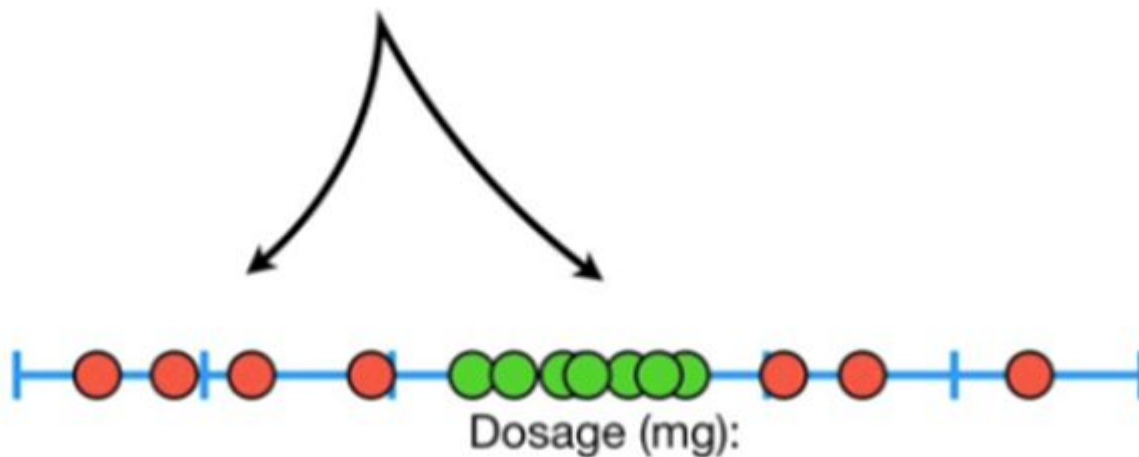
TRANSDUCTIVE SVM (S3VM)

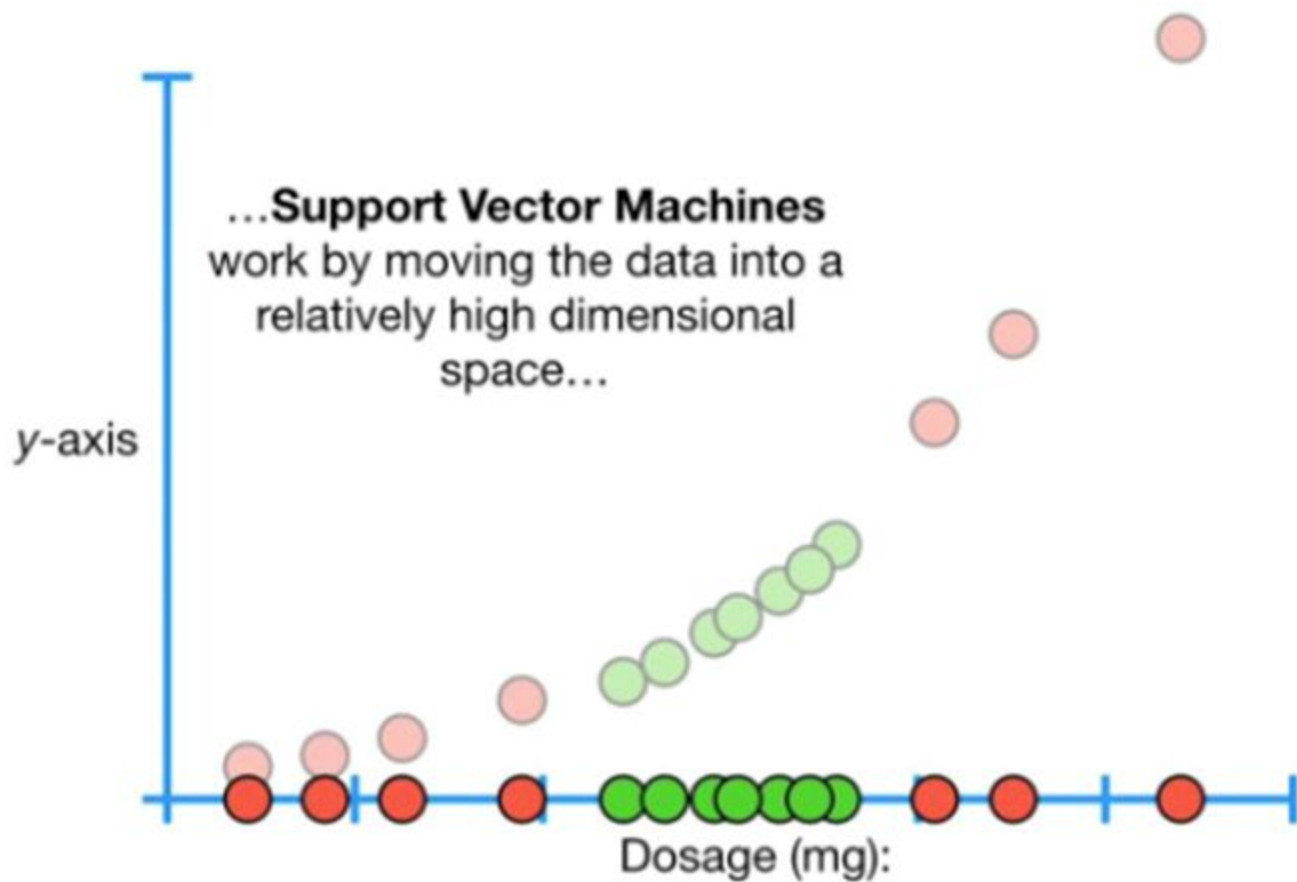
TRANSDUCTION

- Transduction is reasoning from observed, specific(training) cases to specific(test) cases.
- It is used only to test our training data.



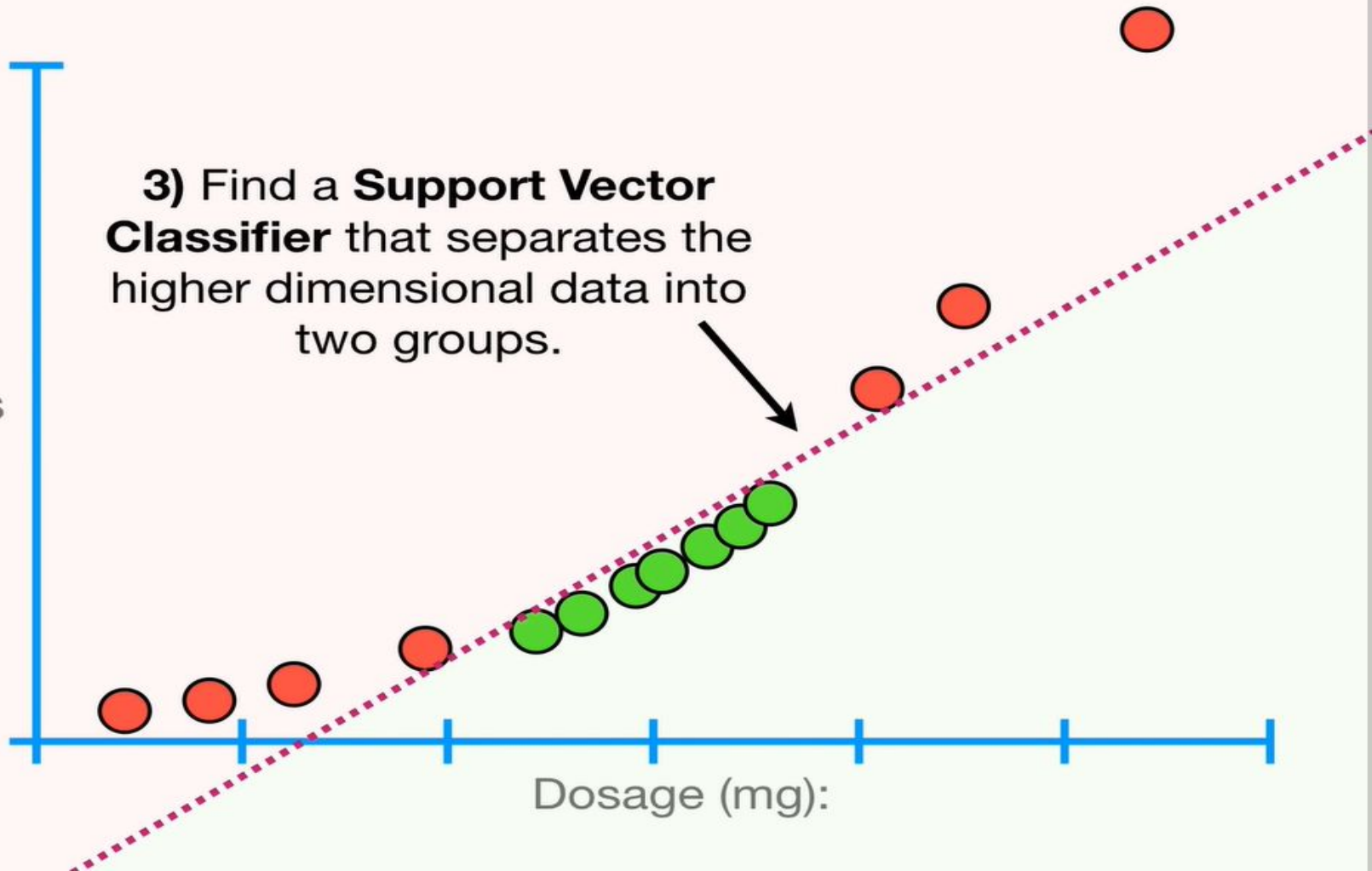
- When we have 2 categories, but no obvious linear classifier that separates them in a nice way...





3) Find a **Support Vector Classifier** that separates the higher dimensional data into two groups.

y-axis



Dosage (mg):

TRANSDUCTIVE SVM (S3VM)

- The Idea: Find largest margin classifier, such that, unlabeled data are outside of the margin as much as possible, use regulation over unlabeled data
- Given Training set $T = \{x_i\}$, and unlabeled set $U = \{u_j\}$
- Find all possible labeling's $U_1 \dots U_n$ on U
- For each $T_k = T \cup U_k$ train a standard SVM
- Choose SVM with largest margins
- NP hard problem, fortunately approximations exist



Thank
you!