

So you want to do supervised learning  
but don't have an intern to label data for  
you....

# Semi-Supervised Learning:

“What is it?

Does it work?

How do you do it?

It doesn't seem like it should work.”

# Basic Setup - Semi-supervised Learning

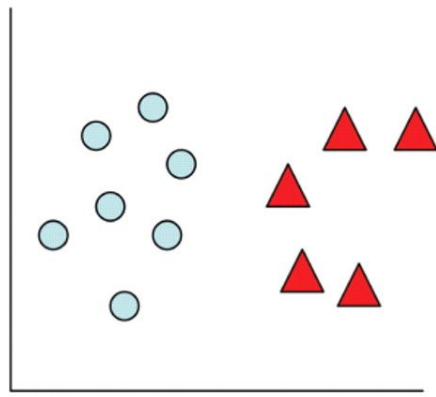
We have a bunch of data. Most of it is unlabeled, some of it is labeled.

We want to take advantage of unlabeled data to build a better classifier than we could by just using the labeled data.



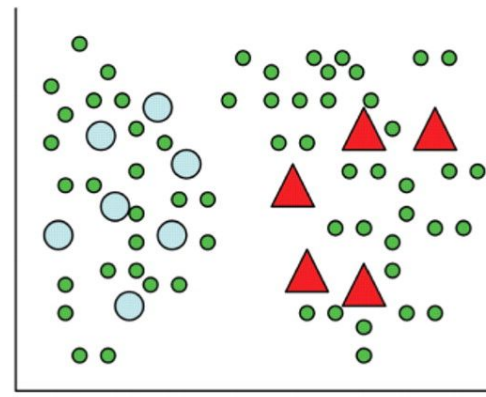
# Objection!

The unlabeled data is unlabeled. How is it going to help us?



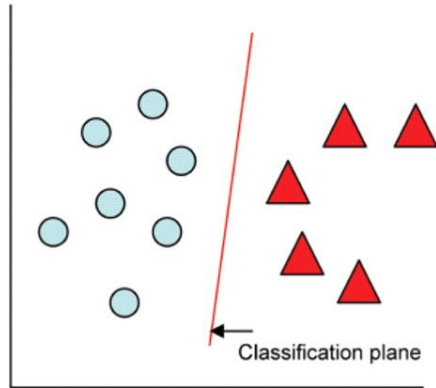
Labeled Data

(a)



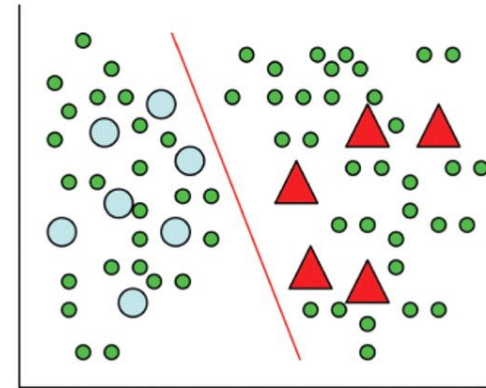
Labeled and Unlabeled Data

(b)



Supervised Learning

(c)



Semi-Supervised Learning

(d)

# Definitions

Semi-supervision

Distant supervision

Weak supervision

Active Learning

Bootstrapping, Self-training


Inductive v. transductive

# Distant Supervision

Distant Supervision: We have labels of some kind, and we have data, and we have some mechanism for mapping those labels onto our data.

Labels are inferred and may be incorrect.

Personal details	
Born	Barack Hussein Obama II August 4, 1961 (age 55) <a href="#">Honolulu, Hawaii, U.S.</a>
Political party	<a href="#">Democratic</a>
Spouse(s)	<a href="#">Michelle Robinson</a> (m. 1992)
Children	Malia Sasha
Parents	<a href="#">Stanley Ann Dunham</a> <a href="#">Barack Obama Sr.</a>
Residence	<a href="#">White House</a>
Alma mater	<a href="#">Occidental College</a> <a href="#">Columbia University</a> <a href="#">Harvard University</a>
Religion	<a href="#">Protestantism</a>
Awards	<a href="#">Nobel Peace Prize</a> (2009)
Signature	
Website	<a href="#">White House</a>  <a href="#">Organizing for Action</a>  <a href="#">Obama Foundation</a> 

**Barack Hussein Obama II** (US  [/bəˈrɑːk huːˈseɪn ʊˈbɑːmə/](#)<sup>[1][2]</sup> born August 4, 1961) is the **44th** and current **President of the United States**. He is the **first African American** to hold the office and the first president born outside the **continental United States**. **Born in Honolulu, Hawaii**, Obama is a graduate of **Columbia University** and **Harvard Law School**, where he was president of the *Harvard Law Review*. He was a **community organizer** in

# Weak Supervision

Often used interchangeably with distant supervision (more in NLP).

Sometimes used to refer to correct but imprecise labels (more in vision community?).



Raw image



Full labeling



Weak labeling



# Active Learning

- Same initial setting as semi-supervised learning.
- Budget to label additional examples
- How to choose which samples to label?

Ideas:

- Label those examples the classifier is least certain about
- Train multiple classifiers, label examples on which they disagree
- Label those examples which would most change the model

Re-active learning - allows for re-labeling of points

# Terminology

Are distant supervision, weak supervision, active learning a subset of semi-supervision?

# More Terminology

“Bootstrapping” = “Self-training” = “Self-teaching”

# Inductive v. Transductive Learning

Inductive - Want to label any new data point we see.

Transductive: Just want to label the data we have.

# Semi-supervised learning as regularization

Setting:

- Data  $X$
- Labelspace  $Y$
- Trying to learn mapping  $f: X \rightarrow Y$
- Want to select  $f^*$ , the best  $f \in F$
- Our unlabeled data induces an implicit ordering of  $f \in F$
- If this ordering ranks  $f^*$  highly, we need less labeled data to learn it