Semi-supervised learning:

Given $D_L = \{(x_i, y_i) ; i = 1, \ldots, L\}$
$D_u = \{x_i ; i = L + 1, \ldots, L + U\}$

Goal: Learn $f \ (y = f(x))$
using both data sets

There are many methods

- "Transductive" SVMs
- Graph-based transduction
- Auto-labeling - TODAY
- Unsupervised feature reduction
  of $D_u$ (e.g. PCA, nonlinear MAPs),
  supervised training on $D_L$
- Supervised feature selection on $D_L$,
  unsupervised clustering on $D_u$
Basic idea of auto-labeling:

Use $D_u$ to initialize, iteratively label $D_u$ and estimate new models.

Variations in data selection? (hard decision)

<table>
<thead>
<tr>
<th>Views</th>
<th>$Y$</th>
<th>$N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Self-training</td>
<td>EM</td>
</tr>
<tr>
<td></td>
<td>Co-training</td>
<td>Co-EM</td>
</tr>
</tbody>
</table>

Self-Training:

0. Train initial model on $D_2$
1. Use $\theta^{(i)}$ to label $D_u$ and add most confident ones to $D_p$
2. Train new model $\theta^{(i+1)}$

Iterate until no more samples can be added.
Problems with self-training:
If initial model is not good, you
- add noise
- don't add much besides central points
⇒ model gets worse

Especially a problem with skewed class distributions
⇒ labeling everything as the majority class makes the problem worse

If initial classifier is very good, then adding unlabeled data doesn't often help

EM: You've seen this in exam estimation problem.
Use self-training with soft labels vs. all data
Taking two views

Co-training
Assumes that the different views are individually enough for classification, conditionally independent given label.
CO-EM: Co-training with soft decisions

Initialize: $\Theta_1^*, \Theta_2^*$ on $D_1$

Iterate:

E-step (A) estimate $p_i^*(y|x; i^*)$

using $\Theta_1^*$

(B) estimate $p_i^*(y|x; i)$

using $\Theta_2^*$

M-step (A) estimate $\Theta_i$ with $p_i^*(y|x; i^*)$

(B) estimate $\Theta_i$ with $p_i^*(y|x; i^*)$

Skewed priors are a problem for co-training like self-training but not for CO-EM

EM & co-EM are better for generative models