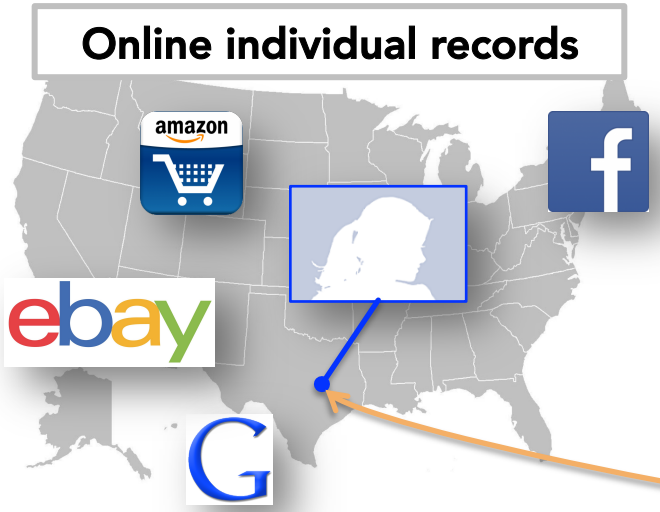
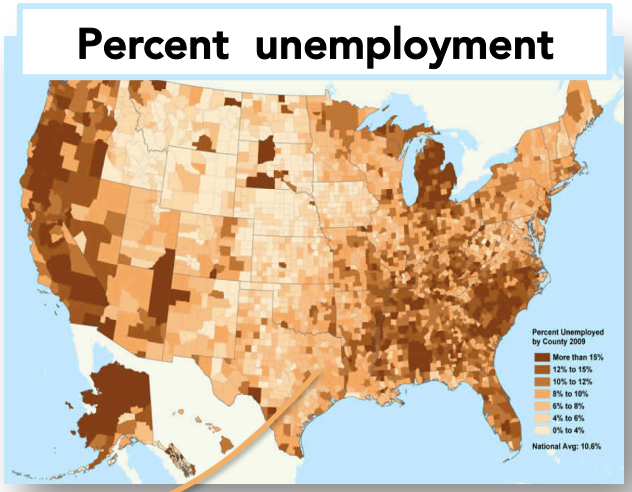
Learning from Label Proportions (LLP)





Input: unlabeled data

Input: label proportions

Output: predictor of individual unemployment

How likely Alice is unemployed
given only her online behavior

Our Solution

Def, Altun&Smola '06: the **mean operator**

$$\mu = 1/m \sum_{i=1}^{m} y_i x_i$$

Thm: μ is **sufficient** for the label variable for most Proper Losses:

$$\text{PROPER-LOSS} = \text{LOSS w/o LABELS}(\theta) - \frac{1}{2} < \theta, \mu >$$

- Quadrianto et al. '09, **homogeneity assumption**:

"Unemployed people in all the counties behave online in the same way"

- Our relaxation:

"The more similar the counties, the more similar the online behavior of the unemployed people"

Results

- **Finite sample approximation bounds** for the resulting classifier (do not hold for previous approaches)
- **First generalization result for LLP**, based on Rademacher complexity