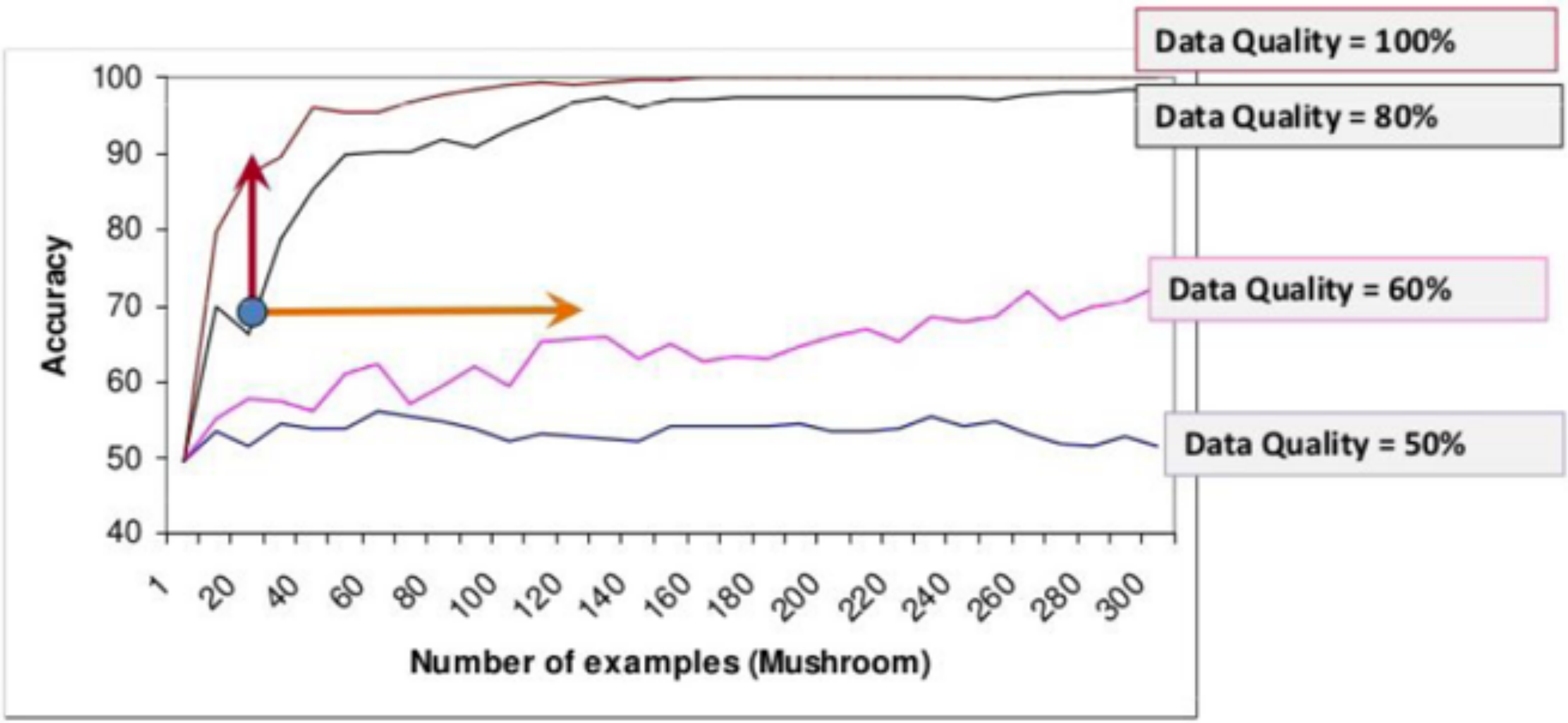
Statistical analysis of MTurk data

- Quality vs. Quantity
- Active Data Collection
- Accounting for worker variation

Different kinds of data

- **Label a word, phrase or pair of texts**
  - "which of the following is a synonym for …"
- **Rate a word, phrase, or pair of texts**
  - "on a scale of 1-5 …
    - how positive is this word?"
    - how similar are these phrases?"
- **Generate or share text**
  - upload an email you wrote
  - translate this sentence
- **Share something about yourself**
  - take a questionnaire and share your tweets

Label Quality vs. Quantity



Ipeirotis, WWW'11

Label Quality vs. Quantity

- **For a fixed budget: trade off number of items rated vs. ratings/item**
- **One rule of thumb:**
  - **With high quality labelers (80% and above):**
    - one label per observation
  - **With low quality labelers (~ 60%):**
    - multiple workers per observation
- **Generally have some overlap to measure inter-annotater agreement**

Sheng et al, KDD 2008,
Kumar and Lease, CSDM 2011

Active Data Collection

- **If first two or three workers agree on an item label, stop collecting for that item**
  - If not get more labels
- **If first workers agrees with predicted label stop collecting for that item**
  - If not get more labels
- **But often not worth it**
  - just throw out labels from low accuracy Turkers

Active Learning

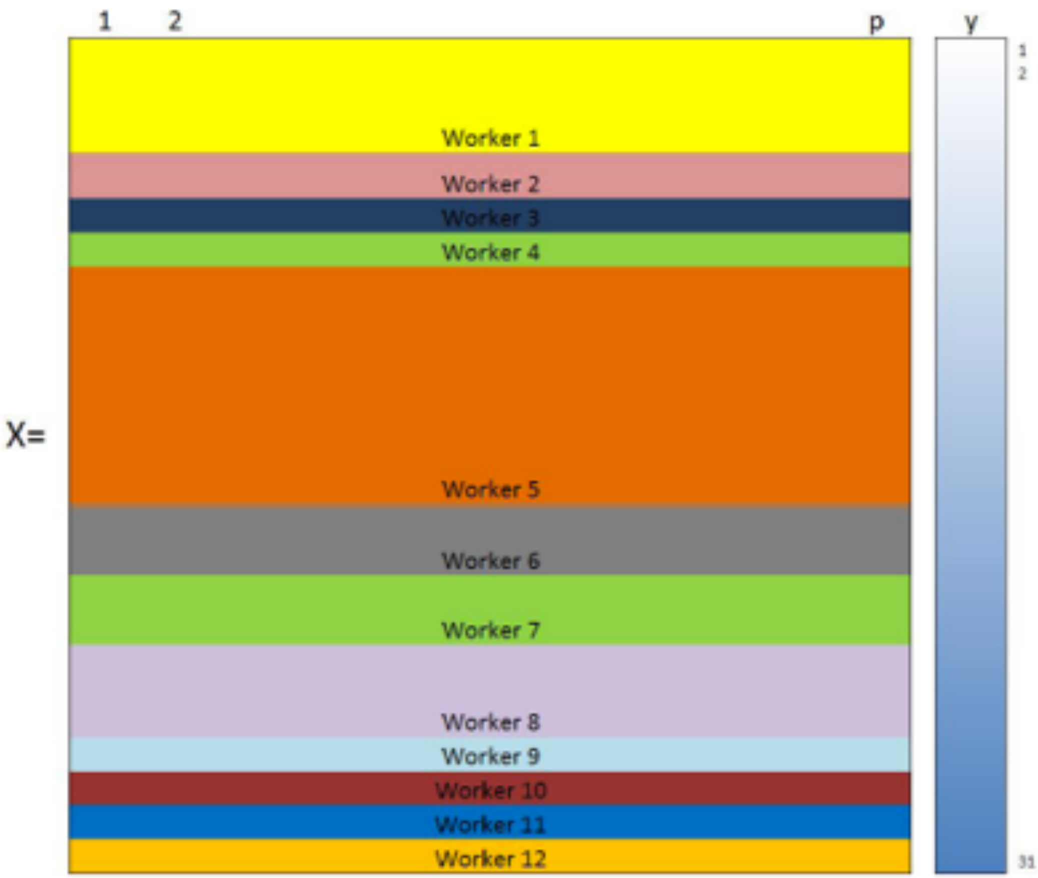
- **Select items to label which are expected to most improve the machine learning model**
  - Items for which the model is uncertain
  - Bayesian estimage of expected change in model
    - *sequential experimental design*

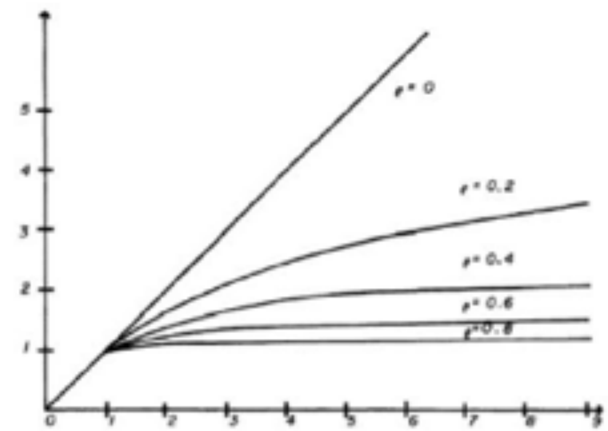
M-turk labeling has block structure

Building $\hat{S}$ requires a machine learning algorithm which usually expects:

$$Y_i = f(x_{1i}, \dots, x_{pi}) + \mathcal{E}_i, \quad \mathcal{E}_1, \dots, \mathcal{E}_n \overset{iid}{\sim} e(0, \sigma^2)$$

What does data from MTurk *really* look like?





Thus, $n_{\text{eff}} \in [12, 31]$, an inconvenient reality that off-the-shelf ML algorithms do not consider:

$$\boldsymbol{\mathcal{E}} \sim e\left(\mathbf{0},\ \sigma^2 \begin{bmatrix} D_1 & & \\ & \ddots & \\ & & D_{12} \end{bmatrix}\right)$$

Item Response Theory

- **People differ in how harshly they rate things**
  - So adjust each Turker's ratings by subtracting off their average rating
- **Problems differ in how hard they are**
  - So give Turkers more credit for getting hard problems right

Using workers of differing quality

- **Use expectation-maximization algorithm:**
  0. Initialize with the aggregate labels being the majority vote
  1. Estimate confusion matrix for each worker
  2. Re-estimate aggregate labels, weighting by worker accuracy (if gold data exists, keep it)
  3. Repeat steps 2-3 until convergence criteria is met

Dawid and Skene (1979)
For a Bayesian version see Raykar et al (2010).

Block structure in practice

- **Do: adjust each Turker's ratings by subtracting off their average**
- **Otherwise: usually ignore correlation but there is vast literature on IRT and on estimating data with block structure**
  - Typically use random-effects linear models or maximum likelihood linear models
  - For classification, use generalized estimating equations (Liang and Zeger, 1986).
  - More recent work uses random forests for cluster-correlated data (Karpievitch et al., 2009) or random-effects expectation-maximization trees (Sela and Simonoff, 2011)

When analyzing people

- **Sometimes want to model people based on their language**
  - Male/Female? Age?
  - Depressed? Extraverted?
  - Liberal/Conservative
- **Want a representative sample**
  - use prescreening questions
  - use Qualtrix or other source of workers
  - re-stratify (reweight) the results.