

Lecture 9-1: Crowdsourcing and Active Learning

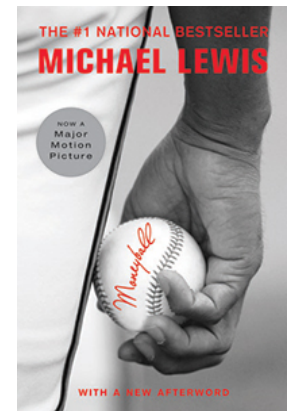
CMPT 733, SPRING 2017

JIANNAN WANG



Data Science Job

Extract value from data



Key Resources

Algorithms

- Machine Learning, Statistical Methods
- Prediction, Business Intelligence

Machines

- Clusters and Clouds
- Warehouse Scale Computing

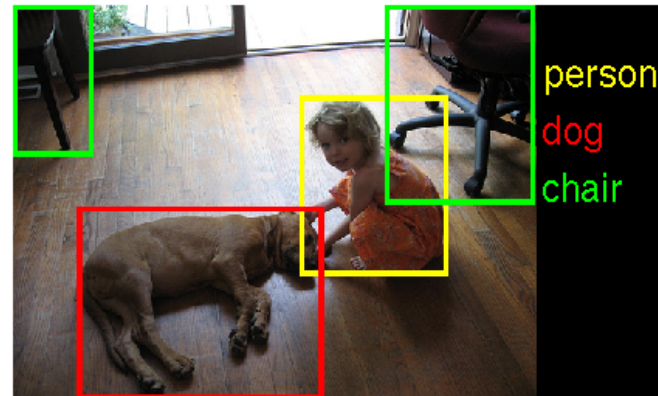
People

- Crowdsourcing, Human Computation
- Data Scientists, Analysts



An Example of Using Three Resources

What are in the image?



How to solve the problem?

Deep Learning (Algorithms)
GPU Cluster (Machines)
ImageNet (People)

What is Crowdsourcing?

Outsourcing

- Allocates work to a **defined** organizational entity

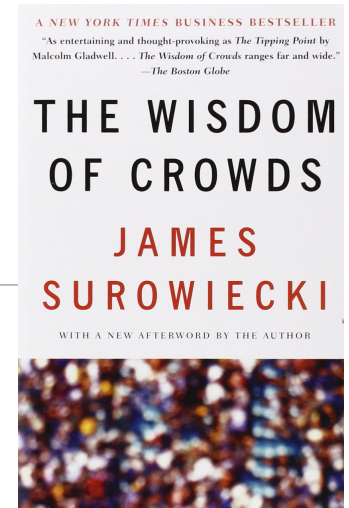
Crowdsourcing

- Allocates work to an **unorganized** collection of individuals

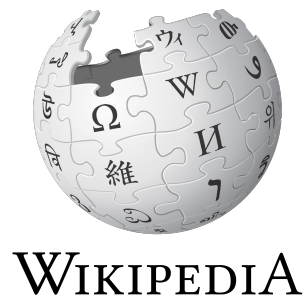
The Wisdom of Crowds

What does it mean?

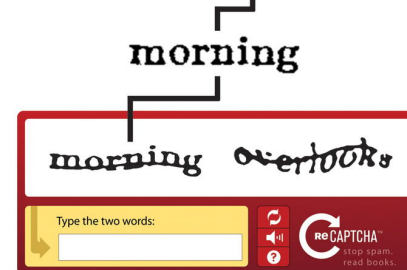
- Two heads are better than one



Some famous examples



The Norwich line steamboat train, from New-London for Boston, this morning ran off the track seven miles north of New-London.



Crowdsourcing Platforms

UpworkTM

mobileworks

amazon **mechanical turk**TM
Artificial Artificial Intelligence

microWorkers
work & earn or offer a micro job

samasource


CrowdFlower

 **minijobz.com**

Amazon Mechanical Turk

500K+ workers*

The screenshot shows the top navigation bar with 'amazonmechanical turk' logo and 'Artificial Intelligence' tagline. It includes links for 'Your Account', 'HITS', and 'Qualifications'. A central banner reads: 'Mechanical Turk is a marketplace for work. We give businesses and developers access to an on-demand, scalable workforce. Workers select from thousands of tasks and work whenever it's convenient. 694,300 HITS available. View them now.'

Make Money by working on HITs

HITs - Human Intelligence Tasks - are individual tasks that you work on. [Find HITs now.](#)

As a Mechanical Turk Worker you:

- Can work from home
- Choose your own work hours
- Get paid for doing good work



[or learn more about being a Worker](#)

Get Results from Mechanical Turk Workers

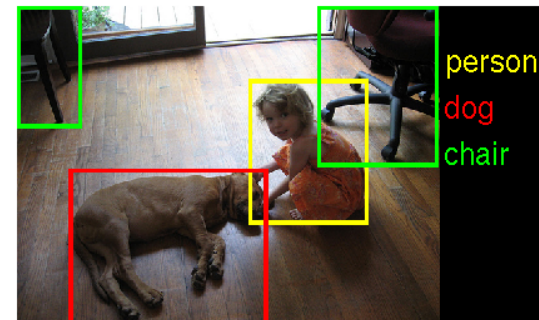
Ask workers to complete HITs - Human Intelligence Tasks - and get results using Mechanical Turk. [Get Started.](#)

As a Mechanical Turk Requester you:

- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you're satisfied with the results



The screenshot shows the requester's dashboard. It includes a search bar with 'HITS' and a filter for 'that pay at least \$ 0.00'. A timer shows '00:00:00 of 2 minutes'. A task is listed: 'Identify if two receipts are the same' by requester 'Jon Erellig' with a reward of '\$0.01 per HIT' and '2 minutes' duration. A 'Total Earned: Unavailable' and 'Total HITs Submitted: 0' are also visible.



* <https://requester.mturk.com/tour>

Crowdsourcing For Data Labeling

Trade-off

- **Cost.** How much will it cost?
- **Latency.** How long will it take?
- **Quality.** How accurate will it be?

Cost Control

- Task Selection, Answer Deduction, Pruning

Latency Control

- Task Pricing, Straggler Mitigation, Pool Maintenance

Quality Control

- Worker Elimination, Answer Aggregation, Task Assignment

Crowdsourcing may not work 😞

What if your data is so big?

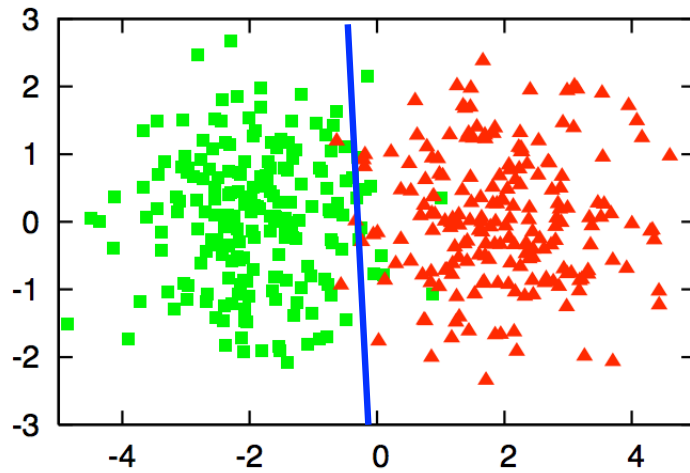
- Label **10 million** images
 - How Long? (1 image / sec) $\frac{10,000,000}{3600*24} = 116 \text{ days}$
 - How much? (\$ 0.1 / image) $10,000,000 * 0.1 = \$1M$

What if your data is confidential?

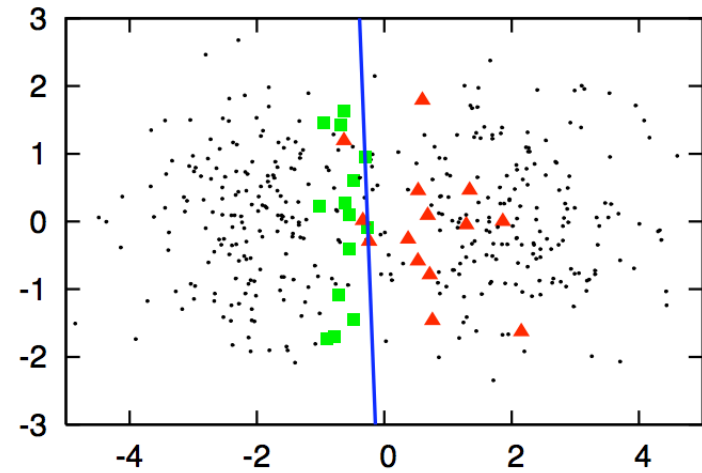
- Medical Data
- Customer Data
- ...

Active Learning

Supervised Learning

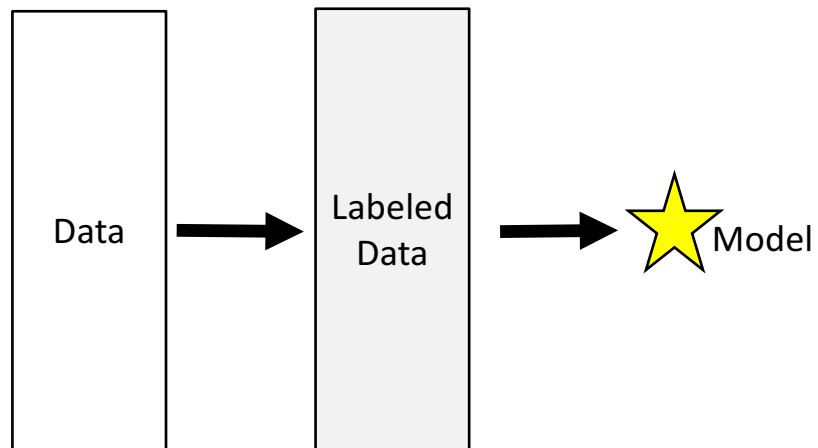


Active Learning

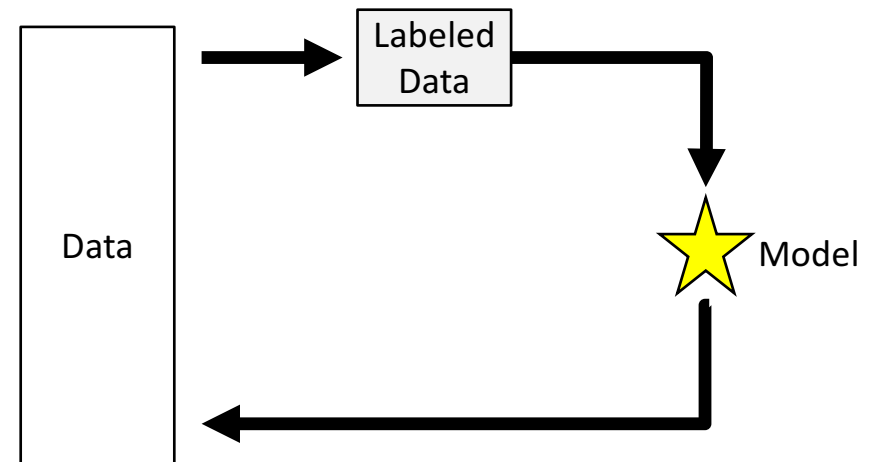


Workflow

Supervised Learning



Active Learning



Query Strategy

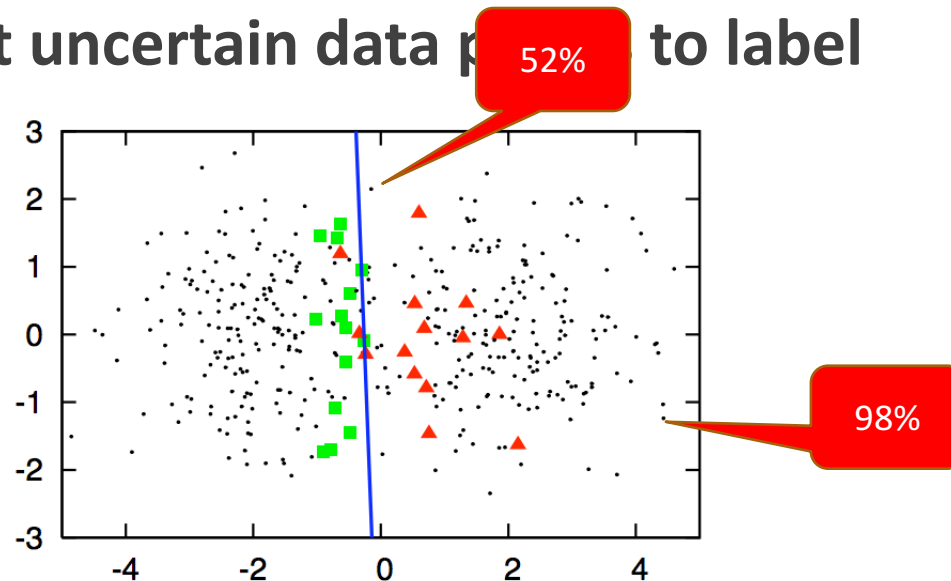
How to decide which data points should be labeled?

- Uncertain Sampling
- Query-By-Committee
- Expected Error Reduction
- Expected Model Change
- Variance Reduction
- Density-Weighted Methods

Settles, Burr. "Active learning literature survey." University of Wisconsin, Madison 52.55-66 (2010): 11.

Uncertain Sampling

Pick up most uncertain data points to label



Logistic Regression

- `predict_proba(X)`

Conclusion

Crowdsourcing

- Why crowdsourcing?
- How does it work?

Active Learning

- Why active learning?
- How does it work?

Assignment 9

Part 1: Data Labeling

- Step 1. Read Data
- Step 2. Removing Obviously Non-matching Pairs
- **Step 3. Active Learning (Task A)**
- Step 4. Model Evaluation

Deadline: 11:59pm, Mar 26th

<http://tiny.cc/cmpt733-a9>