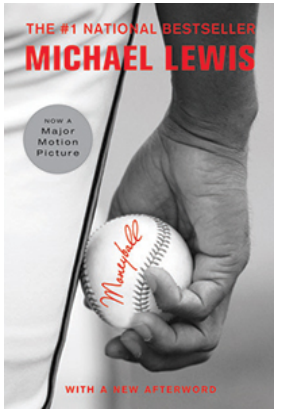
Lecture 9-1:
Crowdsourcing and Active Learning

CMPT 733, SPRING 2017

JIANNAN WANG

Data Science Job

Extract value from data



Key Resources

Algorithms
- Machine Learning, Statistical Methods
- Prediction, Business Intelligence

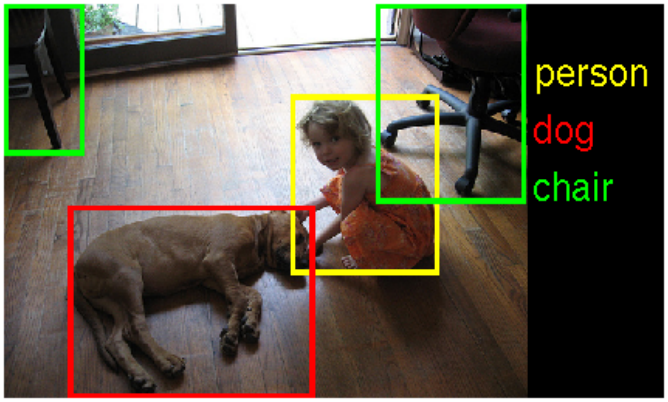
Machines
- Clusters and Clouds
- Warehouse Scale Computing

People
- Crowdsourcing, Human Computation
- Data Scientists, Analysts

From Prof. Michael Franklin's talk

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?

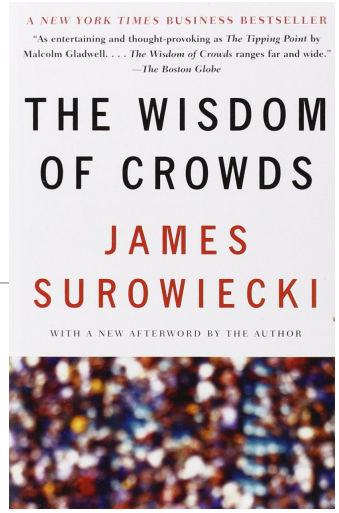
Outsourcing

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

Crowdsourcing

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

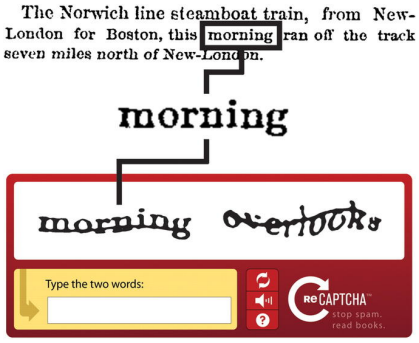
The Wisdom of Crowds



What does it mean?

- Two heads are better than one

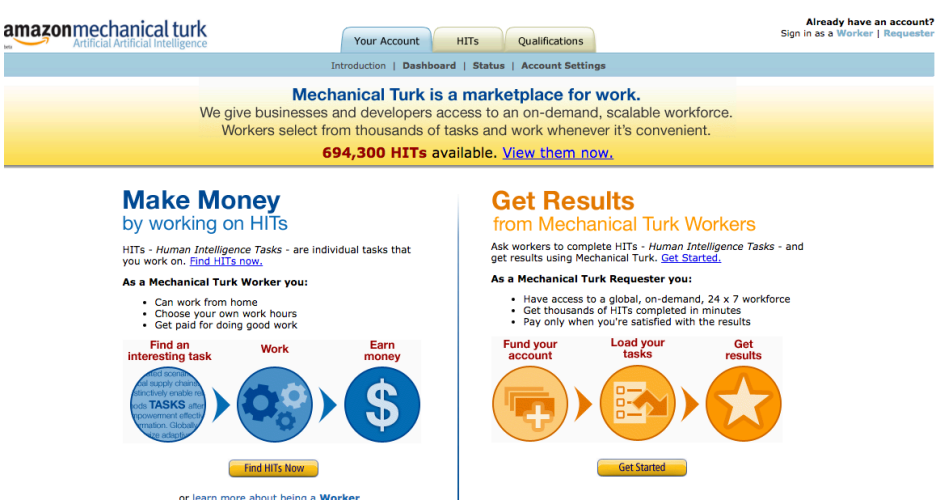
Some famous examples

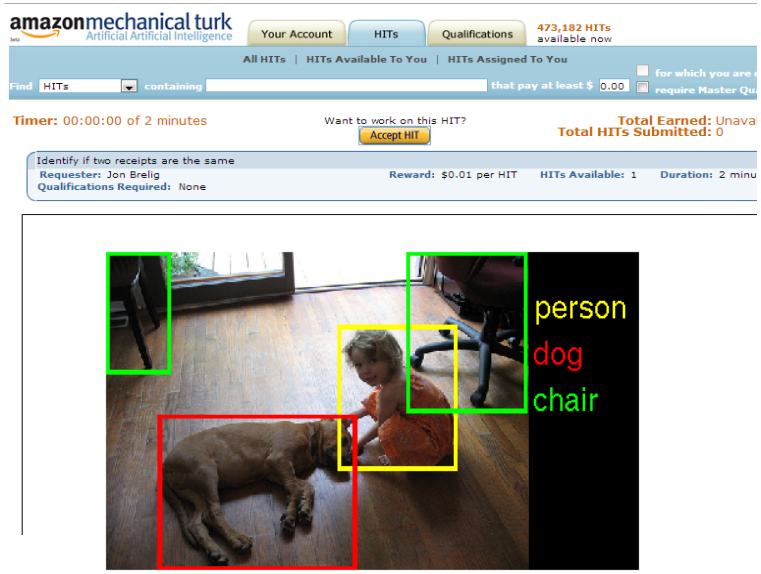


Crowdsourcing Platforms

Amazon Mechanical Turk

500K+ workers*





* 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

1.Guoliang Li, Jiannan Wang, Yudian Zheng, Michael Franklin. “Crowdsourced data management: A survey.” TKDE 2016

Crowdsourcing may not work ☹

What if your data is so big?

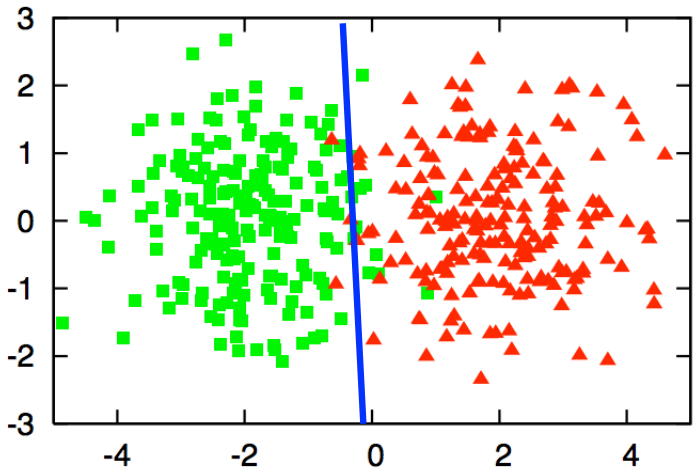
- Label 10 million images
  - How Long? (1 image / sec) $\frac{10{,}000{,}000}{3600*24} = 116$ days
  - How much? ($ 0.1 / image) $10{,}000{,}000 * 0.1 =$ $1M

What if your data is confidential?

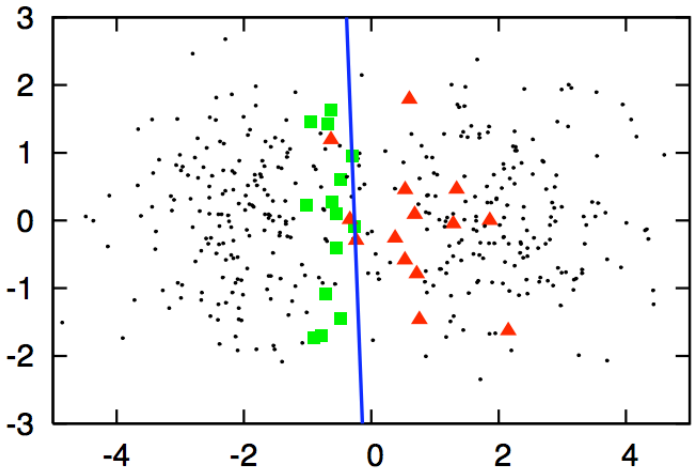
- Medical Data
- Customer Data
- …

Active Learning

Supervised Learning

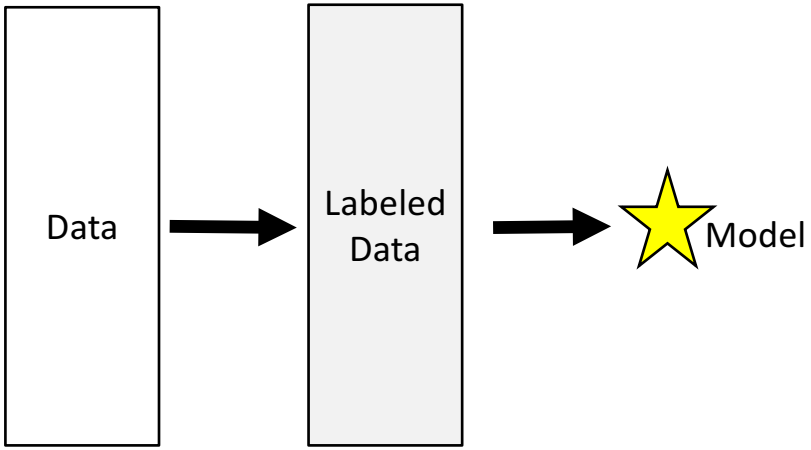


Active Learning

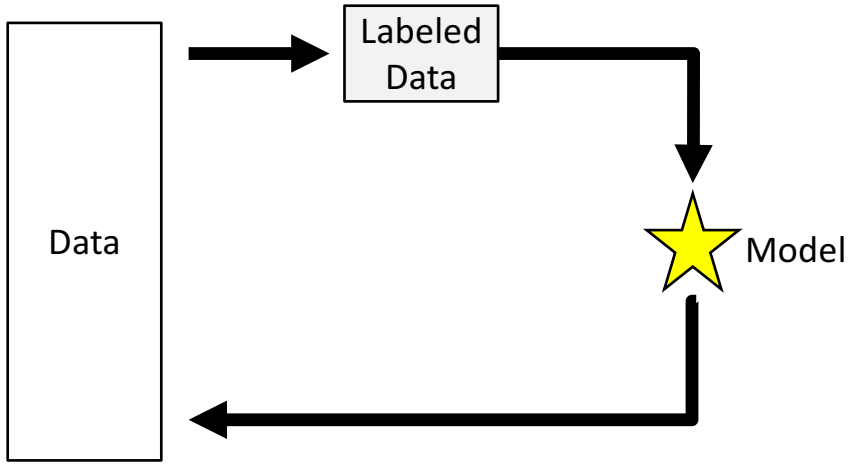


Workflow

Supervised Learning



Active Learning



Query Strategy

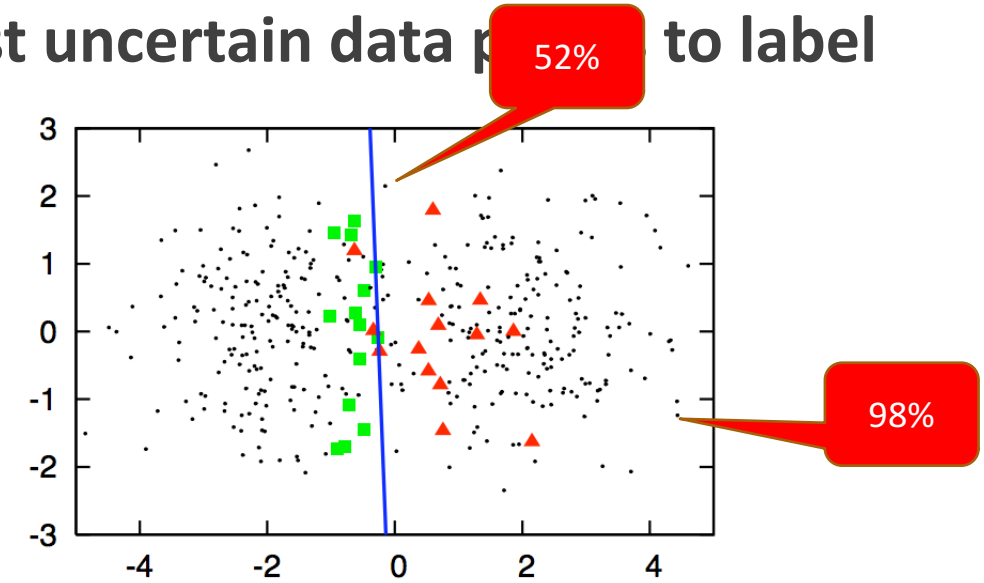
How to decide which data points should be labeled?

- 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 p 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