Label Aggregation: Belief Propagation and Others

Chien-Ju (CJ) Ho

Logistics: Motivation of Assignment 1

- My MTurk (half) Workday. Jeff Bigham.
  - http://www.cs.cmu.edu/~jbigham/posts/2014/half-workday-as-turker.html

- The required reading next lecture also focuses on the worker perspective
  - The methodology/style of the paper is very different from other papers we read

Logistics: Assignment 2

- Will be posted soon-ish.
- Due: October 4 (Friday)

- Programming assignments
  - Implement and compare the performance of majority voting, EM, and SVD
  - You can use any programming languages you like
  - You will be graded based on the report
  - You need to submit your codes
    - Used for plagiarism tests
    - Might check the codes if we have confusions/doubts on the reported results

Logistics: Assignment 2

- Dataset:
  - https://sites.google.com/site/nlpannotations. (rte.standardized.tsv)
  - Recognizing Textual Entailment (RTE) task

Text:

- Many experts think that there is likely to be another terrorist attack on American soil within the next five years.

Hypothesis:

- There will be another terrorist attack on American soil within the next five years.

Answer: NO

Task: Whether the first sentence implies the second hypothesis

- 800 tasks; 164 workers; 10 labels per task

You might want to convert the labels/gold to {+1,-1}

Logistics: Assignment 2

Worker Label

Ground Truth (only used to evaluate the aggregation algorithm)

Worker ID

Task ID

Ignore this column

!amt_annotation_ids	!amt_worker_ids	orig_id	response	gold
89KZPYXSTGTJ0CZY2Y1ZB28YQ9GBT88Z2W1KDYTZ	A19IBSKBTABMR3	266	1	1
89KZPYXSTGTJ0CZY2Y1ZYAJC56Z6FBPGXJYVPXM0	AEX5NCH03LWSG	266	1	1
89KZPYXSTGTJ0CZY2Y1ZFWHATWX49Y3ZTPX4FYH0	A17RPF5ZMO75GW	266	1	1
89KZPYXSTGTJ0CZY2Y1ZV89Z3WRZ6R8ZM4ZZZ070	A15L6WGIK3VU7N	266	0	1
89KZPYXSTGTJ0CZY2Y1ZWZHZYZCCYYVYPDZVNRAZ	A3U7T47F498T1P	266	1	1
89KZPYXSTGTJ0CZY2Y1Z09PZYS137RPZT6SY4A20	AXBQF8RALCIGV	266	1	1
89KZPYXSTGTJ0CZY2Y1ZQ30CJXY2EB96XJS543YZ	A1DCEOFAUIDY58	266	1	1
89KZPYXSTGTJ0CZY2Y1ZXZ3ZNYY7VZKZSCY0B94Z	A1Q4VUJBMY78YR	266	0	1
89KZPYXSTGTJ0CZY2Y1ZDZGGWVYY8XDZTKYC9XKZ	A18941IO2ZZWW6	266	1	1
89KZPYXSTGTJ0CZY2Y1Z3Z7ZWY9J4WFMX60VRVXZ	A11GX90QFWDLMM	266	1	1
89KZPYXSTGTJ0CZY2Y1ZB28YQ9GBT88Z2W1KDYTZ	A19IBSKBTABMR3	934	0	0
89KZPYXSTGTJ0CZY2Y1ZYAJC56Z6FBPGXJYVPXM0	AEX5NCH03LWSG	934	0	0
89KZPYXSTGTJ0CZY2Y1ZFWHATWX49Y3ZTPX4FYH0	A17RPF5ZMO75GW	934	0	0
89KZPYXSTGTJ0CZY2Y1ZV89Z3WRZ6R8ZM4ZZZ070	A15L6WGIK3VU7N	934	0	0
89KZPYXSTGTJ0CZY2Y1ZWZHZYZCCYYVYPDZVNRAZ	A3U7T47F498T1P	934	0	0
89KZPYXSTGTJ0CZY2Y1Z09PZYS137RPZT6SY4A20	AXBQF8RALCIGV	934	1	0
89KZPYXSTGTJ0CZY2Y1ZQ30CJXY2EB96XJS543YZ	A1DCEOFAUIDY58	934	0	0
89KZPYXSTGTJ0CZY2Y1ZXZ3ZNYY7VZKZSCY0B94Z	A1Q4VUJBMY78YR	934	0	0

Logistics: Assignment 2

- Key Requirements
  - Create subsampled datasets:
    - Original: 800 tasks; 164 workers; 10 labels per task
    - Sub-sample the labels, such that each task has k labels
      - k=1, 2, 3, …, 10
  - Implement majority voting, EM (simple one in D&S), and SVD
  - Calculate the error (ratio of tasks the algorithms make wrong predictions)
  - Compare the performance of algorithms
    - Generate a figure with x-axis being k, y-axis being error
    - Plot 3 curves, each corresponding to an algorithm
  - Offer brief discussion

Logistics: Presentations

- Presentation schedule is finalized. Check the course website.

- Talk to me **one week before your presentation**.
  - Default: talk to me after class
  - Finish reading the required paper and at least 1~2 optional papers.
  - Have an idea of your presentation plan.
  - Come up with at least two discussion questions.
  - Try to look for materials outside of the reading list to include in your presentation (e.g., real-world examples/platforms related to the topics you present)

Logistics: Project Proposal

- Due this Friday.

- If you want to chat this week, you can schedule a meeting:
  - https://calendly.com/chienjuho/30min

- I'll also reserve time after the proposal deadline.

Logistics: Announcements

- Please constantly check these places for announcements
  - Course website
  - Piazza
  - The first few slides of each lecture
    - Especially if you are not in class or late to class

What We Learned So Far

- EM-based methods
  - Empirically performs well
  - Relatively computationally efficient
  - No theoretical guarantee

- Matrix-based methods
  - Comes with theoretical guarantee
  - Computationally expensive
  - Require some “potentially unreasonable” assumptions for the analysis

This Lecture

- More label aggregation methods
  - Iterative message passing
  - Minimax entropy
  - Variational inference

- Discussion on common assumptions and limitations

- Examples on human-in-the-loop learning

Iterative Learning for Reliable Crowdsourcing Systems

Karger, Oh, and Shah. NIPS 2011.

Iterative Message Passing

- One of the more well-cited papers in label aggregation
  - One of the early non-EM-based papers
  - The algorithm is very simple and intuitive
  - Solid theoretical guarantees
  - One of the first to formally address the task assignment question

Setting (Basically the same as the paper in our previous lecture)

- Basic components
  - m tasks $i = 1, \dots, m$
  - n workers $j = 1, \dots, n$
  - $A_{i,j} \in \{+1, -1\}$: label provided by worker j to task i

- Label generation process
  - Each task i has a true label $s_i \in \{+1, -1\}$
  - Each worker j gives correct label with probability p_j

$$\mathbf{A}_{ij} = \begin{cases} s_i & \text{with probability } p_j \,, \\ -s_i & \text{with probability } 1 - p_j \,, \end{cases}$$

Essentially
- Homogeneous tasks
- Unknow worker skills

Key Idea of the Proposed Algorithm

	Task 1	Task 2	Task 3	Task 4	…
Worker 1	1	-1	1	1	
Worker 2	1	-1	-1	-1	
Worker 3	-1	1	-1	1	
Worker 4	1	-1	1	1	
…					

Labels are in $\{1, -1\}$

One-min short discussion:

What do you think is the most likely label for task 3? Why?

(Think about EM we have discussed before)

EM Performs Iterative Updates

	Task 1	Task 2	Task 3	Task 4	…	skill estimates	
Worker 1	1	-1	1	1		0.875	1
Worker 2	1	-1	-1	-1		0.625	0.5
Worker 3	-1	1	-1	1		0.375	0.25
Worker 4	1	-1	1	1		0.875	1
…							
label estimates	1	-1	?	1			
	1	-1	1	1			

Note:
The algorithm in the paper shares similar ideas but is not exactly the same.

Unlikely to Have a Complete Matrix

- In the previous lecture
  - Assume each rater i has a probability p_i to rate a contribution

$$u_{ti} = \begin{cases} q_t & \text{w.p. } p_i\psi_i \\ -q_t & \text{w.p. } p_i(1-\psi_i) \\ 0 & \text{w.p. } 1-p_i. \end{cases}$$

Not an ideal model.
However, it makes the analysis tractable.

- In this work, assume the requester can ***assign tasks*** to workers
  - Want to achieve similar property above (in a more deterministic way)

Task Allocation via Regular Graph

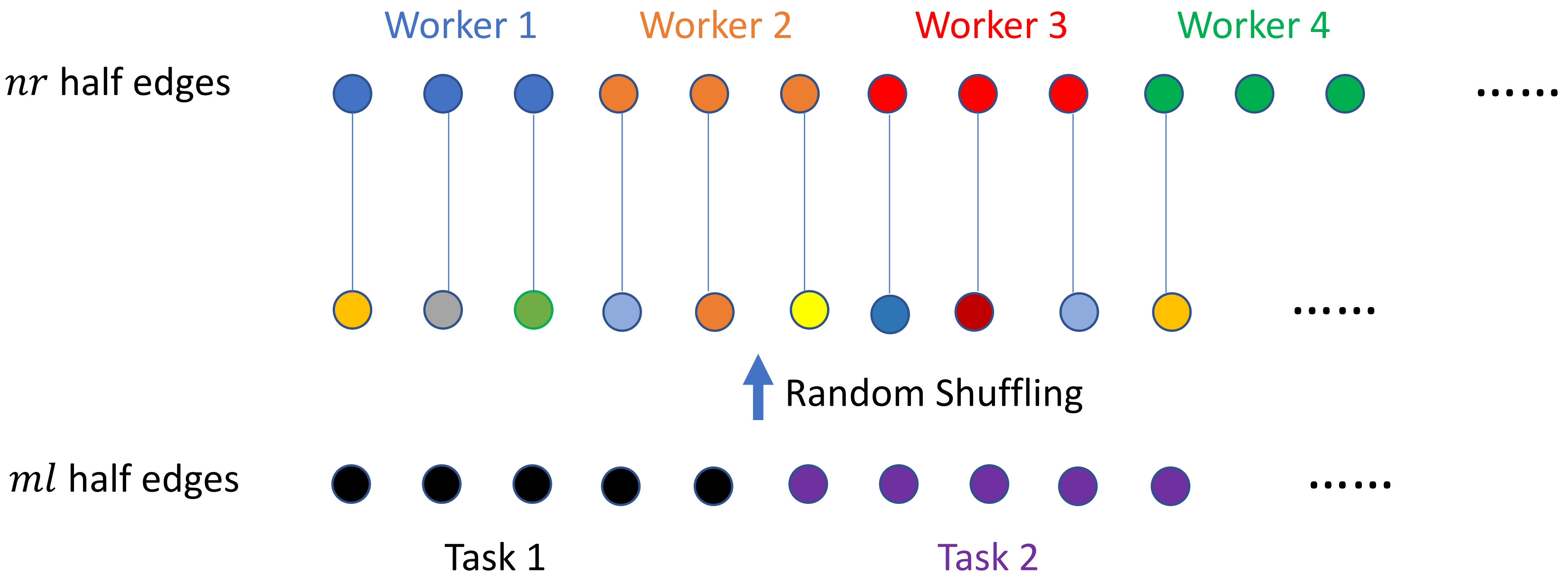
- Requirements:
  - m tasks; n workers
  - Each worker is assigned r tasks
  - Each task is assigned to l workers
  - $ml = nr$

- Assume $m \gg r$ and $n \gg l$

- Goal: Try to make the allocation as "uniform" as possible

(l, r)-Regular Bipartite Graph

- An informal description
  - Each worker is assigned $r = 3$ tasks
  - Each task is assigned to $l = 5$ workers
  - How to make the assignments as "uniform" as possible?



Inference via Iterative Message Passing

- Notations
  - Task message at round k: $x_{i\to j}^{(k)}$
  - Worker message at round k: $y_{j\to i}^{(k)}$

Task messages:
How likely the true label is +1

Worker messages:
How likely the worker gives correct labels?

- Algorithm:
  - Randomly initialize worker message from a normal distribution
  - For $k = 1$ to k_{max}

$$
\begin{aligned}
x_{i\to j}^{(k)} &= \sum_{j'\in\partial i\backslash j} A_{ij'} y_{j'\to i}^{(k-1)} , \\
y_{j\to i}^{(k)} &= \sum_{i'\in\partial j\backslash i} A_{i'j} x_{i'\to j}^{(k)} ,
\end{aligned}
$$

∂_j: the set of tasks assigned to worker j
∂_i : the set of workers who are assigned task i
A : label matrix

  - Prediction according to $sign(\sum_{j\in\partial_i} A_{ij} y_{j\to i}^{(k_{max})})$

Key Intuitions in the Analysis

- Using approximated tree structure
  - (l, r)-regular bipartite graph gives a local tree structure in message passing
  - Each leaf is drawn from Gaussian distribution
  - The root is a *weighted sum of Gaussian variables*

- Key techniques used
  - Estimate the mean and variance of the weighted sum of Gaussian variables
  - Using the estimated mean and variance to prove the tail bounds

Main Results (1)

Corollary 2.2. *Under the hypotheses of Theorem 2.1, there exists* $m_0 = 3\ell r e^{\ell q/4\sigma_\infty^2}(\hat{\ell}\hat{r})^{2(k-1)}$ *and* $k_0 = 1 + \left(\log(q/\mu^2)/\log(\hat{\ell}\hat{r}q^2)\right)$ *such that*

$$\frac{1}{m}\sum_{i=1}^{m}\mathbb{P}\left(t_i \neq \hat{t}_i^{(k)}\right) \leq 2e^{-\ell q/(4\sigma_\infty^2)}, \tag{4}$$

for all $m \geq m_0$ *and* $k \geq k_0$.

- Expected error = $O(e^{-lq})$
  - l: number of workers per task
  - $q = E[(2p_j - 1)^2]$ is the expected “quality” of the workers

Main Results (2)

Theorem 2.7. *When $q \leq C$ for any constant $C < 1$, there exists a positive constant C' such that*

$$\min_{\tau \in \widetilde{\mathcal{T}}_\ell, \hat{t}} \max_{t \in \{\pm 1\}^m, \mathcal{F} \in \mathcal{F}_q} \frac{1}{m} \sum_{i \in [m]} \mathbb{P}(t_i \neq \hat{t}_i) \geq \frac{1}{2} e^{-C' \ell q}, \tag{5}$$

for all m where the task assignment scheme τ ranges over all adaptive schemes that use at most $m\ell$ queries and $\hat{t}$ ranges over all estimators that are measurable functions over the responses.

- In this setting, no matter what task assignment algorithm you use (even adaptive ones), the expected error is at least $\Omega(e^{-lq})$

A very strong claim, but what is the claim really saying?

Discussion

- What do you think about the paper?
- When do you think task assignment helps? Why task assignment is not helpful in their setting?

Task Assignment

- We had a lecture for task assignment last semester, but I removed it this semester due to the packed schedule.

Optimization: Task Assignment	**Required** Online Task Assignment in Crowdsourcing Markets. Ho and Vaughan. AAAI 2012. **Optional** Pick-a-Crowd: Tell Me What You Like, and I'll Tell You What To Do. Difallah et al. WWW 2013. Using Hierarchical Skills for Optimized Task Assignment in Knowledge-Intensive Crowdsourcing. Mavridis et al. WWW 2016. Adaptive Task Assignment for Crowdsourced Classification. Ho, Jabbari, and Vaughan. ICML 2013.

Learning from the Wisdom of Crowd by Minimax Entropy

Zhou et al. NIPS 2012.

Entropy (Information Entropy)

- Consider a random variable X with n possible values
- The probability for each value i happening is P_i

- Information entropy (Shannon entropy)

$$H(X) = -\sum_{i=1}^{n} P_i \ln P_i$$

What are the interpretations of entropy?

Higher entropy => More uncertainty => Higher unpredictability

Principle of Maximum Entropy

"the probability distribution which best represents the current state of knowledge is the one with largest entropy"

- Consider a dice with 6 faces
  - Without any knowledge, what's your best bet on the probability of 1~6 happening
  - Assume you are told the probability of 3 happening is ½, what's your best bet on the probability of the rest numbers happening?

Setting

Goal: Given $\vec{z}$, how to infer $\vec{\pi}$ and $\vec{y}$?

Observations

	Task 1	Task 2	Task 3	…	Task n
Worker 1	$\vec{z}_{1,1}$	$\vec{z}_{1,2}$	$\vec{z}_{1,3}$	…	$\vec{z}_{1,n}$
Worker 2	$\vec{z}_{2,1}$	$\vec{z}_{2,2}$	$\vec{z}_{2,3}$	…	$\vec{z}_{2,n}$
Worker 3	$\vec{z}_{3,1}$	$\vec{z}_{3,2}$	$\vec{z}_{3,3}$	…	$\vec{z}_{3,n}$
…	…	…	…	…	…
Worker m	$\vec{z}_{m,1}$	$\vec{z}_{m,2}$	$\vec{z}_{m,3}$	…	$\vec{z}_{m,n}$

Underlying distribution

	Task 1	Task 2	Task 3	…	Task n
Worker 1	$\vec{\pi}_{1,1}$	$\vec{\pi}_{1,2}$	$\vec{\pi}_{1,3}$	…	$\vec{\pi}_{1,n}$
Worker 2	$\vec{\pi}_{2,1}$	$\vec{\pi}_{2,2}$	$\vec{\pi}_{2,3}$	…	$\vec{\pi}_{2,n}$
Worker 3	$\vec{\pi}_{3,1}$	$\vec{\pi}_{3,2}$	$\vec{\pi}_{3,3}$	…	$\vec{\pi}_{3,n}$
…	…	…	…	…	…
Worker m	$\vec{\pi}_{m,1}$	$\vec{\pi}_{m,2}$	$\vec{\pi}_{m,3}$	…	$\vec{\pi}_{m,n}$

- Components
  - Workers $i = 1, \dots, m$
  - Tasks $j = 1, \dots, n$
  - Labels $k = 1, \dots, c$
- Worker labels $\vec{z}_{i,j} = (z_{i,j,1}, \dots, z_{i,j,c})$
  - $z_{i,j,k} = 1$ if worker i label task j as class k
  - $z_{i,j,k} = 0$ otherwise
- True labels $\vec{y}_j = (y_{i,1}, \dots, y_{j,c})$
  - $y_{j,l} = 1$ if task j's label is l
  - $y_{j,l} = 0$ otherwise
- Worker skills: $\vec{\pi}_{i,j} = (\pi_{i,j,1}, \dots, \pi_{i,j,c})$
  - $z_{i,j,k}$: probability for worker i label task j as class k

Apply the Maximum Entropy Principle

- Assume true labels $\vec{y}_j$ are given, how to infer $\vec{\pi}$?
- Choose $\vec{\pi}$ that maximizes entropy subject to the observations of $\vec{z}$

- Choose $\vec{\pi}$ that maximizes entropy subject to the observations of $\vec{z}$

$$\max_{\pi} \quad \boxed{-\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{c} \pi_{ijk} \ln \pi_{ijk}} \quad \textbf{Entropy}$$

s.t.

$$\boxed{\sum_{k=1}^{c} \pi_{ijk} = 1, \ \forall i, j, \ \pi_{ijk} \geq 0, \ \forall i, j, k.} \quad \textbf{Probability constraints}$$

	Task 1	Task 2	Task 3	…	Task n
Worker 1	$\vec{z}_{1,1}$	$\vec{z}_{1,2}$	$\vec{z}_{1,3}$	…	$\vec{z}_{1,n}$
Worker 2	$\vec{z}_{2,1}$	$\vec{z}_{2,2}$	$\vec{z}_{2,3}$	…	$\vec{z}_{2,n}$
Worker 3	$\vec{z}_{3,1}$	$\vec{z}_{3,2}$	$\vec{z}_{3,3}$	…	$\vec{z}_{3,n}$
…	…	…	…	…	…
Worker m	$\vec{z}_{m,1}$	$\vec{z}_{m,2}$	$\vec{z}_{m,3}$	…	$\vec{z}_{m,n}$

	Task 1	Task 2	Task 3	…	Task n
Worker 1	$\vec{\pi}_{1,1}$	$\vec{\pi}_{1,2}$	$\vec{\pi}_{1,3}$	…	$\vec{\pi}_{1,n}$
Worker 2	$\vec{\pi}_{2,1}$	$\vec{\pi}_{2,2}$	$\vec{\pi}_{2,3}$	…	$\vec{\pi}_{2,n}$
Worker 3	$\vec{\pi}_{3,1}$	$\vec{\pi}_{3,2}$	$\vec{\pi}_{3,3}$	…	$\vec{\pi}_{3,n}$
…	…	…	…	…	…
Worker m	$\vec{\pi}_{m,1}$	$\vec{\pi}_{m,2}$	$\vec{\pi}_{m,3}$	…	$\vec{\pi}_{m,n}$

- Choose $\vec{\pi}$ that maximizes entropy subject to the observations of $\vec{z}$

$$\max_{\pi} \quad -\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{c} \pi_{ijk} \ln \pi_{ijk}$$

$$\text{s.t.} \quad \boxed{\sum_{i=1}^{m} \pi_{ijk} = \sum_{i=1}^{m} z_{ijk}, \ \forall j, k,}$$

"expected # labels in each class = empirical # labels

(We can relax them to "approximately equal")

$$\sum_{k=1}^{c} \pi_{ijk} = 1, \ \forall i, j, \ \pi_{ijk} \geq 0, \ \forall i, j, k.$$

	Task 1	Task 2	Task 3	…	Task n
Worker 1	$\vec{z}_{1,1}$	$\vec{z}_{1,2}$	$\vec{z}_{1,3}$	…	$\vec{z}_{1,n}$
Worker 2	$\vec{z}_{2,1}$	$\vec{z}_{2,2}$	$\vec{z}_{2,3}$	…	$\vec{z}_{2,n}$
Worker 3	$\vec{z}_{3,1}$	$\vec{z}_{3,2}$	$\vec{z}_{3,3}$	…	$\vec{z}_{3,n}$
…	…	…	…	…	…
Worker m	$\vec{z}_{m,1}$	$\vec{z}_{m,2}$	$\vec{z}_{m,3}$	…	$\vec{z}_{m,n}$

	Task 1	Task 2	Task 3	…	Task n
Worker 1	$\vec{\pi}_{1,1}$	$\vec{\pi}_{1,2}$	$\vec{\pi}_{1,3}$	…	$\vec{\pi}_{1,n}$
Worker 2	$\vec{\pi}_{2,1}$	$\vec{\pi}_{2,2}$	$\vec{\pi}_{2,3}$	…	$\vec{\pi}_{2,n}$
Worker 3	$\vec{\pi}_{3,1}$	$\vec{\pi}_{3,2}$	$\vec{\pi}_{3,3}$	…	$\vec{\pi}_{3,n}$
…	…	…	…	…	…
Worker m	$\vec{\pi}_{m,1}$	$\vec{\pi}_{m,2}$	$\vec{\pi}_{m,3}$	…	$\vec{\pi}_{m,n}$

- Choose $\vec{\pi}$ that maximizes entropy subject to the observations of $\vec{z}$

$$\max_{\pi} \quad -\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{c}\pi_{ijk}\ln\pi_{ijk}$$

$$\text{s.t.} \quad \sum_{i=1}^{m}\pi_{ijk} = \sum_{i=1}^{m} z_{ijk},\ \forall j,k,\ \boxed{\sum_{j=1}^{n} y_{jl}\pi_{ijk} = \sum_{j=1}^{n} y_{jl} z_{ijk},\ \forall i,k,l,}$$

$$\sum_{k=1}^{c}\pi_{ijk} = 1,\ \forall i,j,\ \pi_{ijk} \geq 0,\ \forall i,j,k.$$

	Task 1	Task 2	Task 3	…	Task n
Worker 1	$\vec{z}_{1,1}$	$\vec{z}_{1,2}$	$\vec{z}_{1,3}$	…	$\vec{z}_{1,n}$
Worker 2	$\vec{z}_{2,1}$	$\vec{z}_{2,2}$	$\vec{z}_{2,3}$	…	$\vec{z}_{2,n}$
Worker 3	$\vec{z}_{3,1}$	$\vec{z}_{3,2}$	$\vec{z}_{3,3}$	…	$\vec{z}_{3,n}$
…	…	…	…	…	…
Worker m	$\vec{z}_{m,1}$	$\vec{z}_{m,2}$	$\vec{z}_{m,3}$	…	$\vec{z}_{m,n}$

	Task 1	Task 2	Task 3	…	Task n
Worker 1	$\vec{\pi}_{1,1}$	$\vec{\pi}_{1,2}$	$\vec{\pi}_{1,3}$	…	$\vec{\pi}_{1,n}$
Worker 2	$\vec{\pi}_{2,1}$	$\vec{\pi}_{2,2}$	$\vec{\pi}_{2,3}$	…	$\vec{\pi}_{2,n}$
Worker 3	$\vec{\pi}_{3,1}$	$\vec{\pi}_{3,2}$	$\vec{\pi}_{3,3}$	…	$\vec{\pi}_{3,n}$
…	…	…	…	…	…
Worker m	$\vec{\pi}_{m,1}$	$\vec{\pi}_{m,2}$	$\vec{\pi}_{m,3}$	…	$\vec{\pi}_{m,n}$

Solving the Optimization

- Given true labels y, we use maximum entropy to find π

=> For every set of true labels y, we obtain π and the corresponding entropy

- How to decide the true labels y?
  - Higher entropy => higher uncertainty
  - Choosing labels that minimize uncertainty/entropy

- Minimax entropy

$$\min_{y} \max_{\pi} \quad -\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{c} \pi_{ijk} \ln \pi_{ijk}$$

$$\text{s.t.} \quad \sum_{i=1}^{m} \pi_{ijk} = \sum_{i=1}^{m} z_{ijk},\ \forall j,k,\ \sum_{j=1}^{n} y_{jl}\pi_{ijk} = \sum_{j=1}^{n} y_{jl} z_{ijk},\ \forall i,k,l,$$

$$\sum_{k=1}^{c} \pi_{ijk} = 1,\ \forall i,j,\ \pi_{ijk} \geq 0,\ \forall i,j,k,\ \sum_{l=1}^{c} y_{jl} = 1,\ \forall j,\ y_{jl} \geq 0,\ \forall j,l.$$

An interesting way of looking at label aggregation

- Finding the labels/distribution with minimax entropy

- Can we incorporate models of label generation?
    - e.g., Tasks are homogeneous
    - e.g., Tasks have different difficulty levels

- Express them as additional constraints

Additional Details on the Technical Insights

- The dual formulation gives nice insights
  - One set of dual variables represent worker skills
  - Another set of dual variable represent task difficulties

- Perform reasonably well in practice

Method	Dogs	Web
Minimax Entropy	84.63	88.05
Dawid & Skene	84.14	83.98
Majority Voting	82.09	73.07
Average Worker	70.60	37.05

A Recap on Label Aggregation

The Approaches We Covered

- EM-Based methods (The mainstream approach)
  - Develop models of label generation
  - Write down the likelihood function
  - Using EM algorithms to optimize likelihood
- Matrix-based method
  - Perform SVD, using the top left singular vector as the prediction
- Others
  - Iterative message passing
    - Be careful about message normalization if you want to use this for assignment 3
  - Minimax entropy

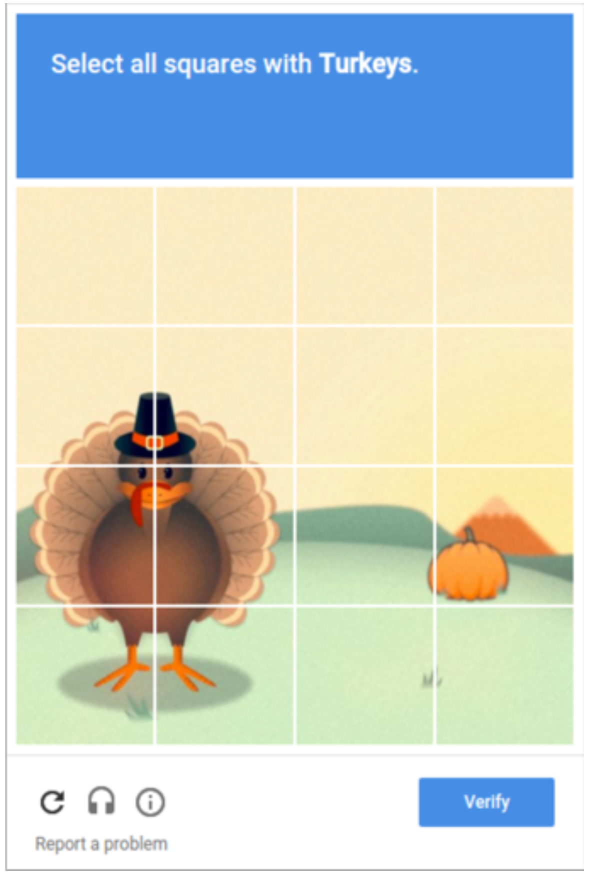
General Discussion on Label Aggregation

- Common assumption: each label is i.i.d. drawn from some distribution
- This assumption enables tons of papers applying statistics/learning techniques in crowdsourcing (low-hanging fruit)

- Discussion
  - What other assumptions have been made in the papers you read?
  - Under what scenarios do you think this (and/or other assumptions) is reasonable?
  - Is there any assumption you think we should try to relax in this line of research.
  - If you need to keep working on label aggregation, what would you propose to do?

Beyond Label Aggregation

Course Overview



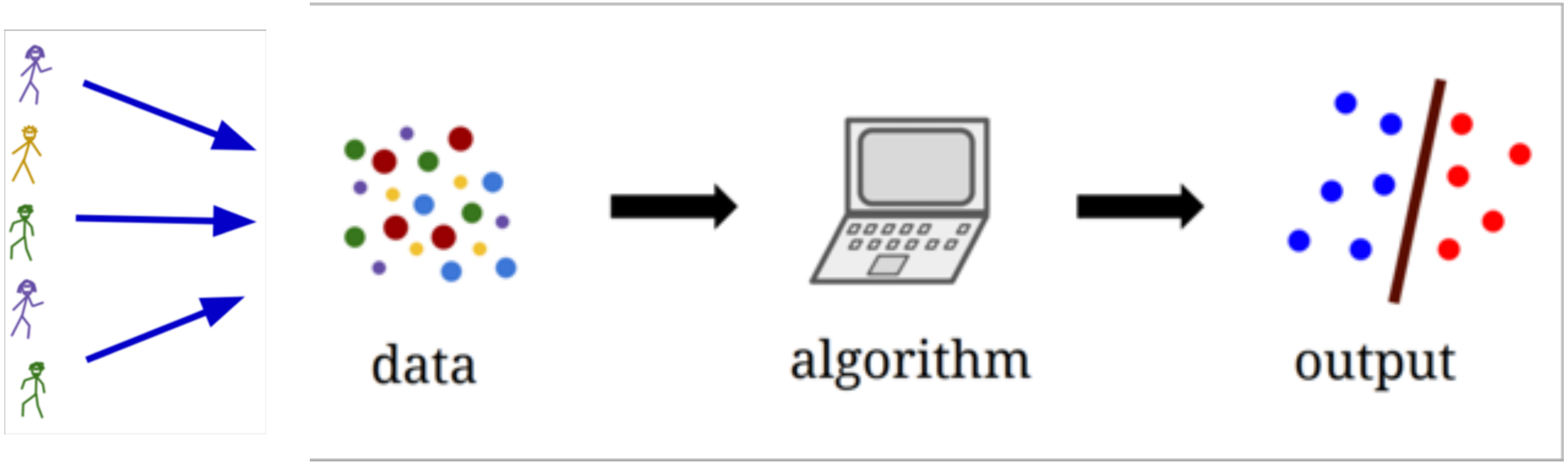
Human as data sources:
Label aggregation
Probabilistic reasoning to aggregate noisy human data

Practical challenges:
Real-time and complex tasks
Studies on workflow and team designs from HCI perspective

Humans are "Humans":
Incentive design
Game theoretical modeling of humans and incentive design

Selected recent topics:
Ethical issues of AI/ML, learning with strategic behavior, Human-AI collaborations.

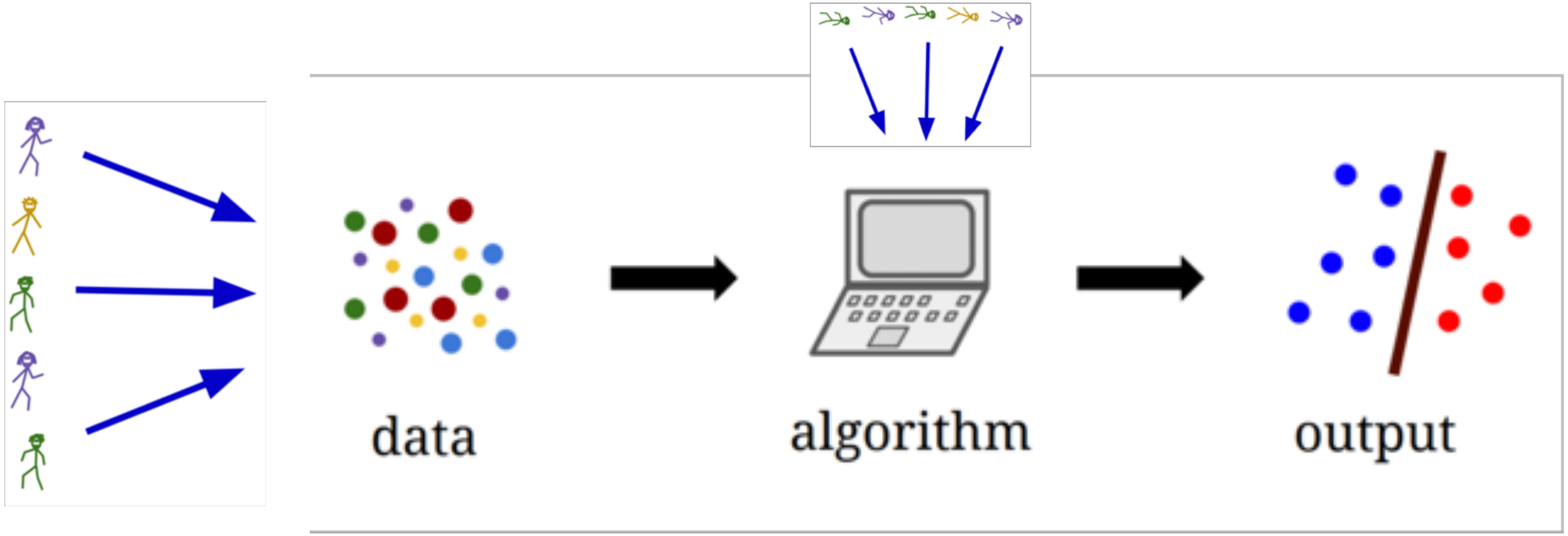
Learning with Humans in the Loop



We have focused on the aspect of “data generated by humans”
Make strong i.i.d. assumptions on the data generation process

1. What if data is not i.i.d.?

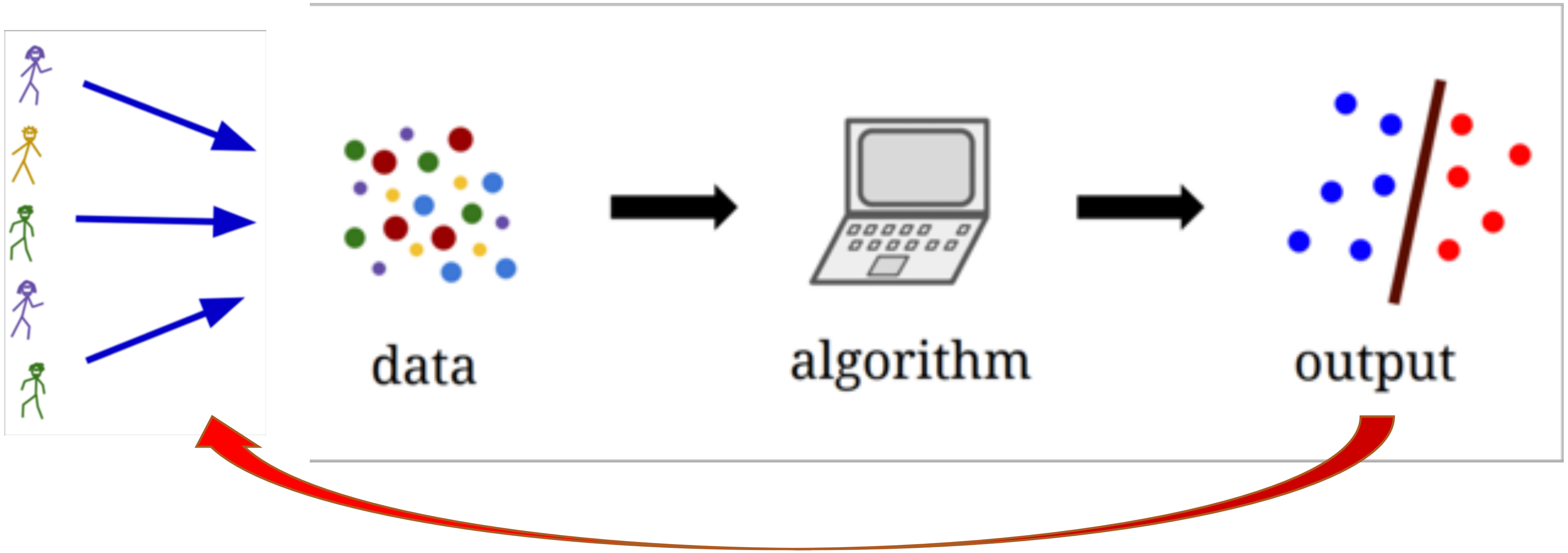
Learning with Humans in the Loop



We have focused on the aspect of “data generated by humans”
Make strong i.i.d. assumptions on the data generation process

2. Can humans be involved in the algorithm as well?

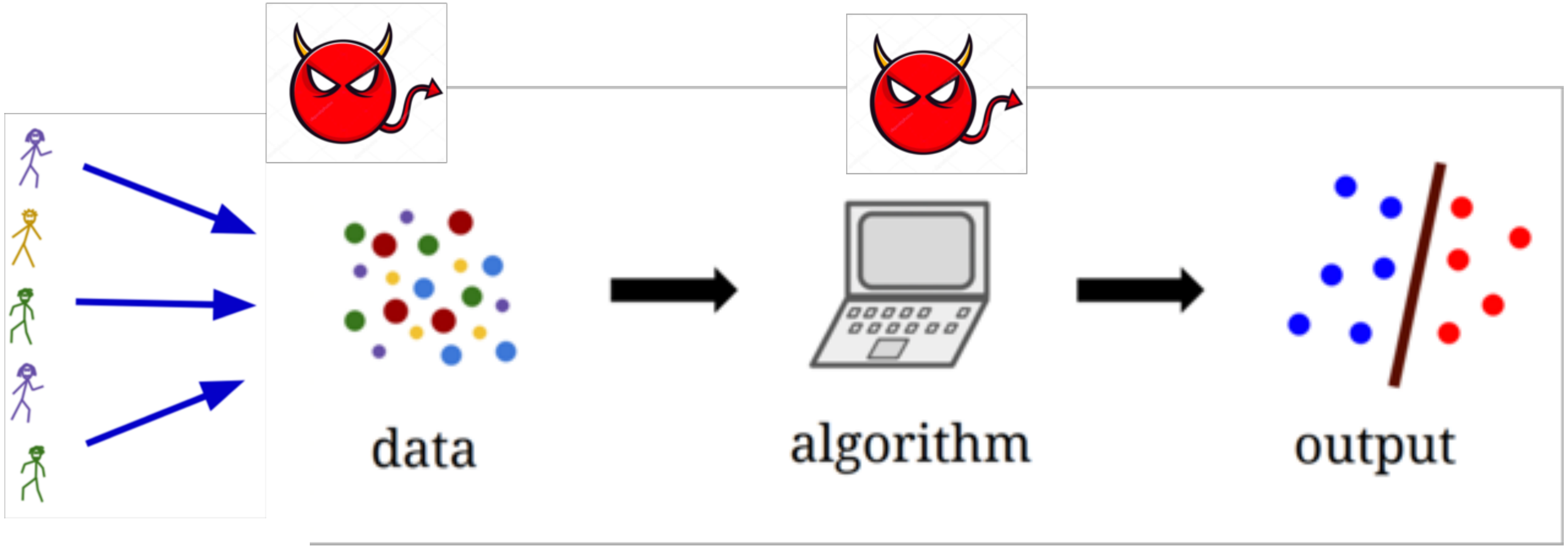
Learning with Humans in the Loop



We have focused on the aspect of "data generated by humans"
Make strong i.i.d. assumptions on the data generation process

3. What if the output has impacts to humans?

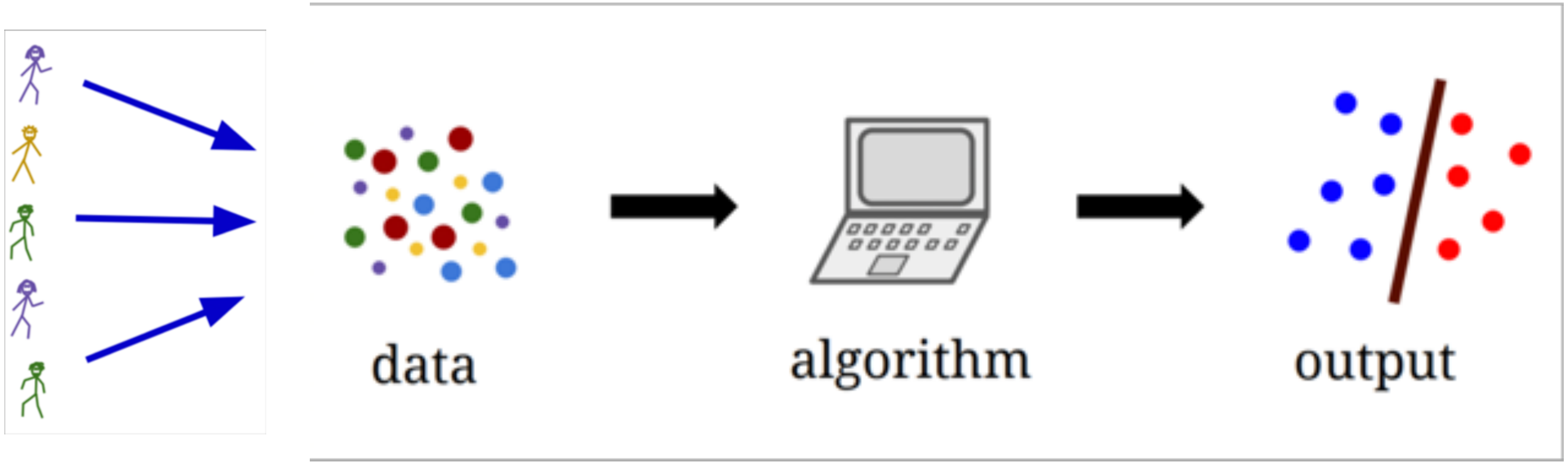
Learning with Humans in the Loop



We have focused on the aspect of “data generated by humans”
Make strong i.i.d. assumptions on the data generation process

4. What if some users try to sabotage the system.

Learning with Humans in the Loop



We have focused on the aspect of “data generated by humans”
Make strong i.i.d. assumptions on the data generation process

1. What if data is not i.i.d.?

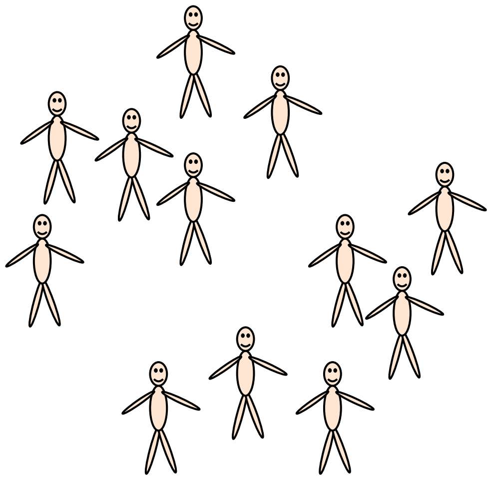
Active Buying Data for Machine Learning

Joint work with
Jake Abernethy, Yiling Chen, and Bo Waggoner

ACM Economics and Computation 2015

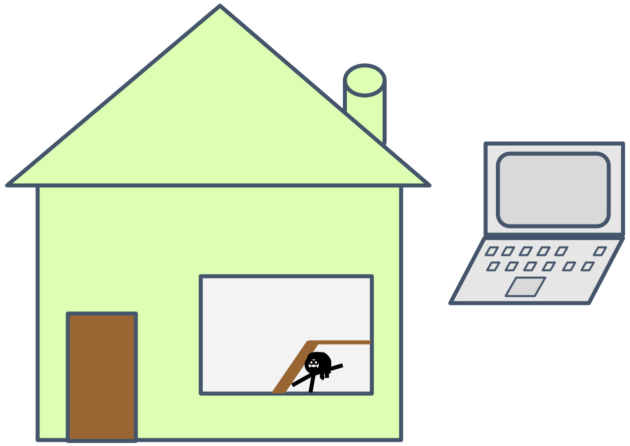
Learning via Buying Data from Humans

data-holders



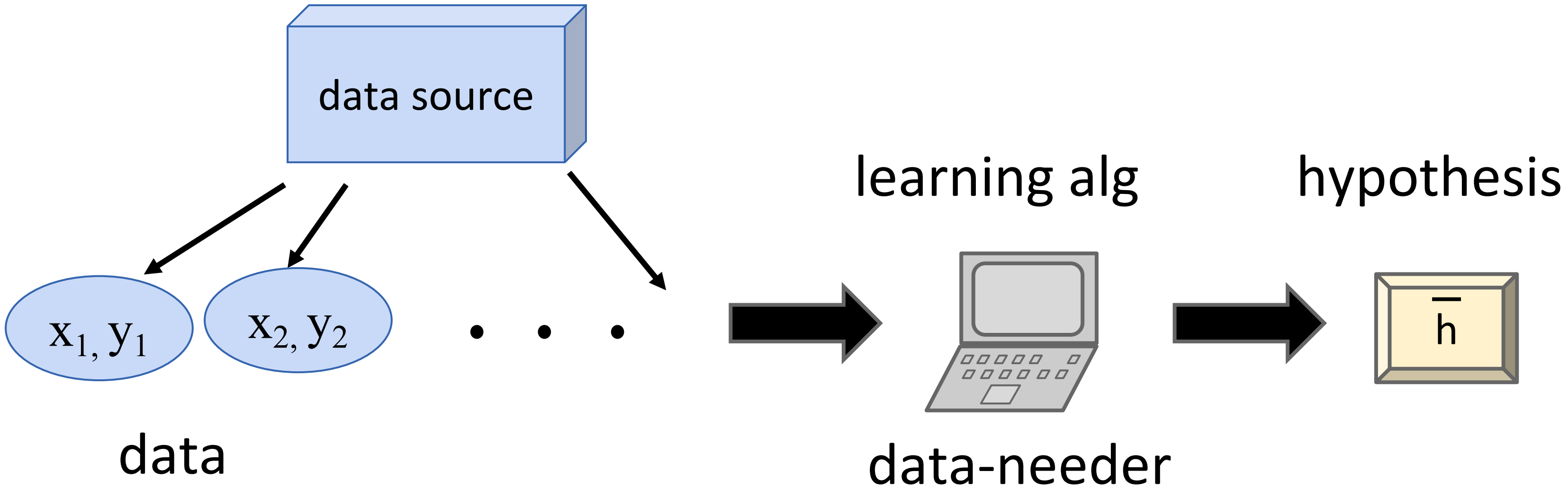
ex: medical data

data-needers

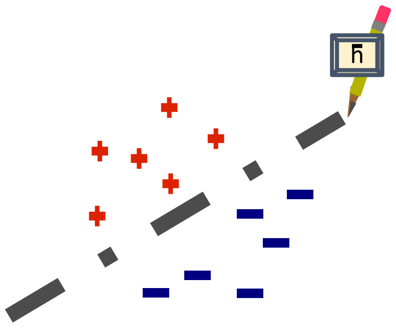


ex: pharmaceutical co.

(Traditional) Learning Problems

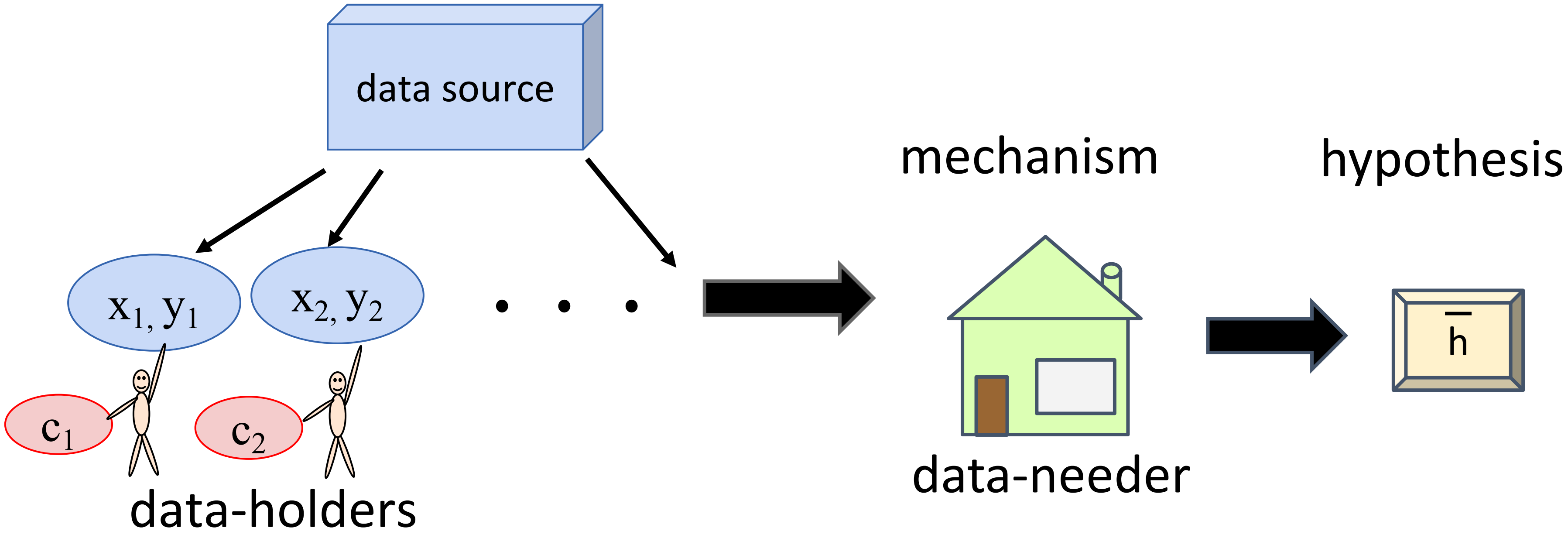


Goal:
Learn a **good** hypothesis *h* with **few** data points



Example: Classification
Data: (point, label) where label is + or −
Hypothesis: hyperplane separating the two types

Our Setting: Data are Held by Humans



Goal:

Learn a **good** hypothesis *h* with **small** budgets

Assumptions:

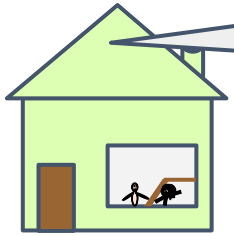
data cannot be fabricated

costs are **unknown** to the data-needer and **bounded**

costs can be arbitrarily **correlated** with data

What can we do?

Want to learn a classifier for HIV
(the maximum cost is $1000)



Paying $1000 for everyone
until the budget is exhausted.

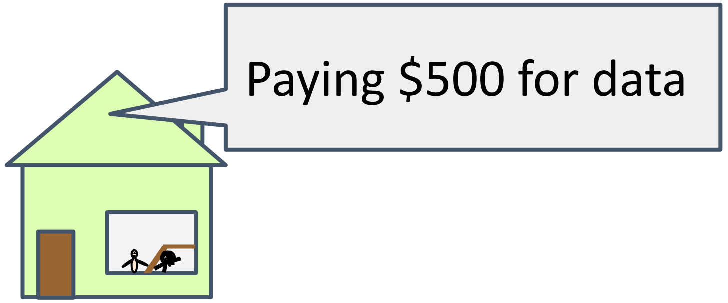
Pro: We can apply standard learning algorithms

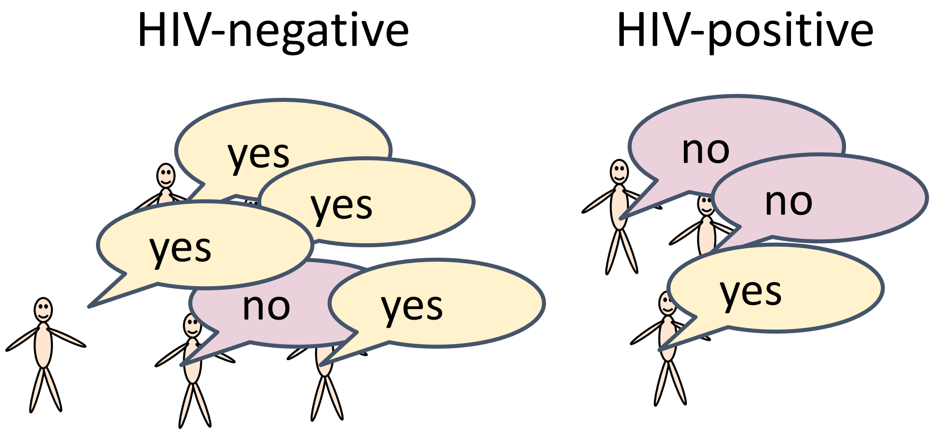
Con: Waste a lot of money

What can we do?

Want to learn a classifier for HIV

(the maximum cost is $1000)

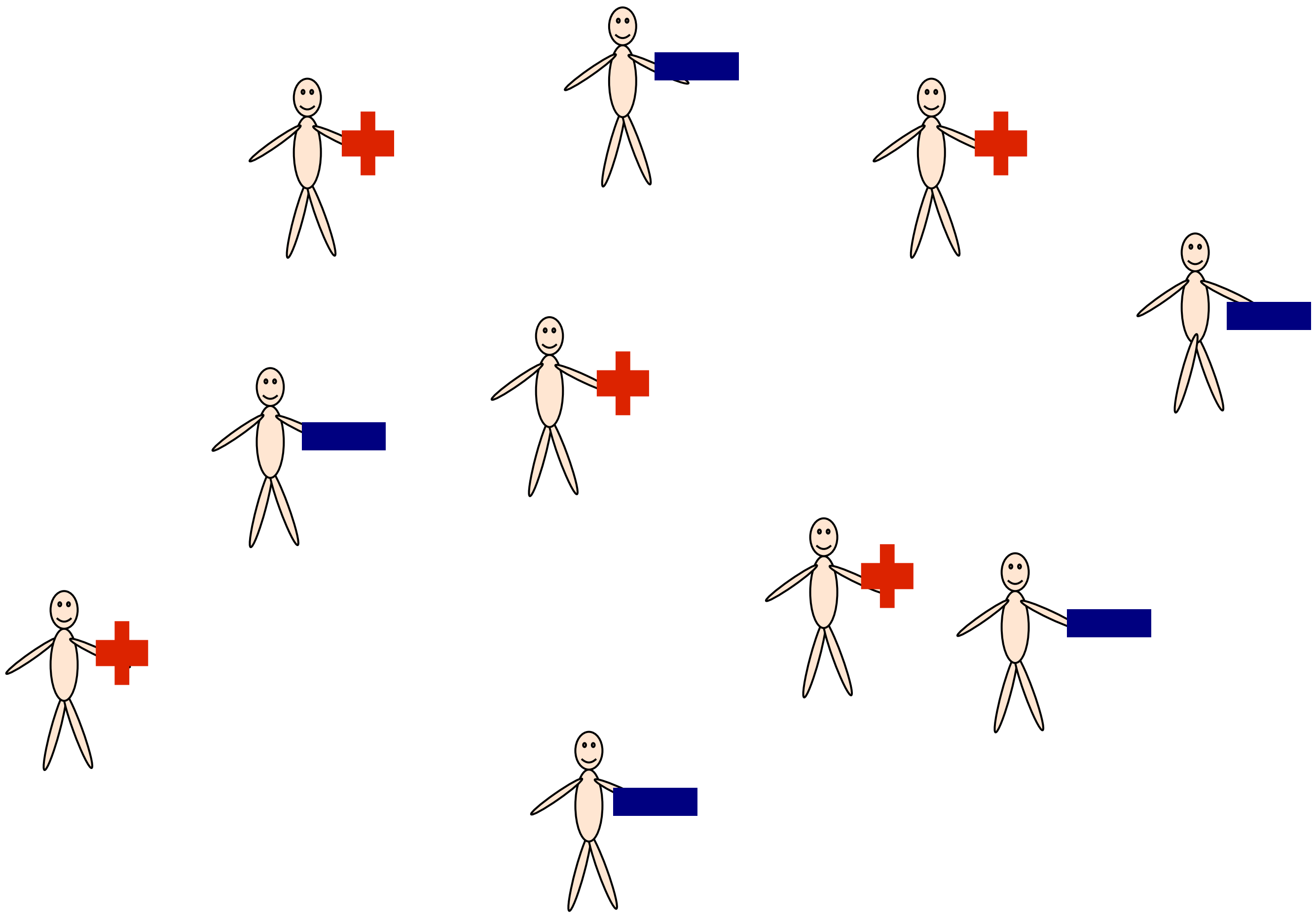




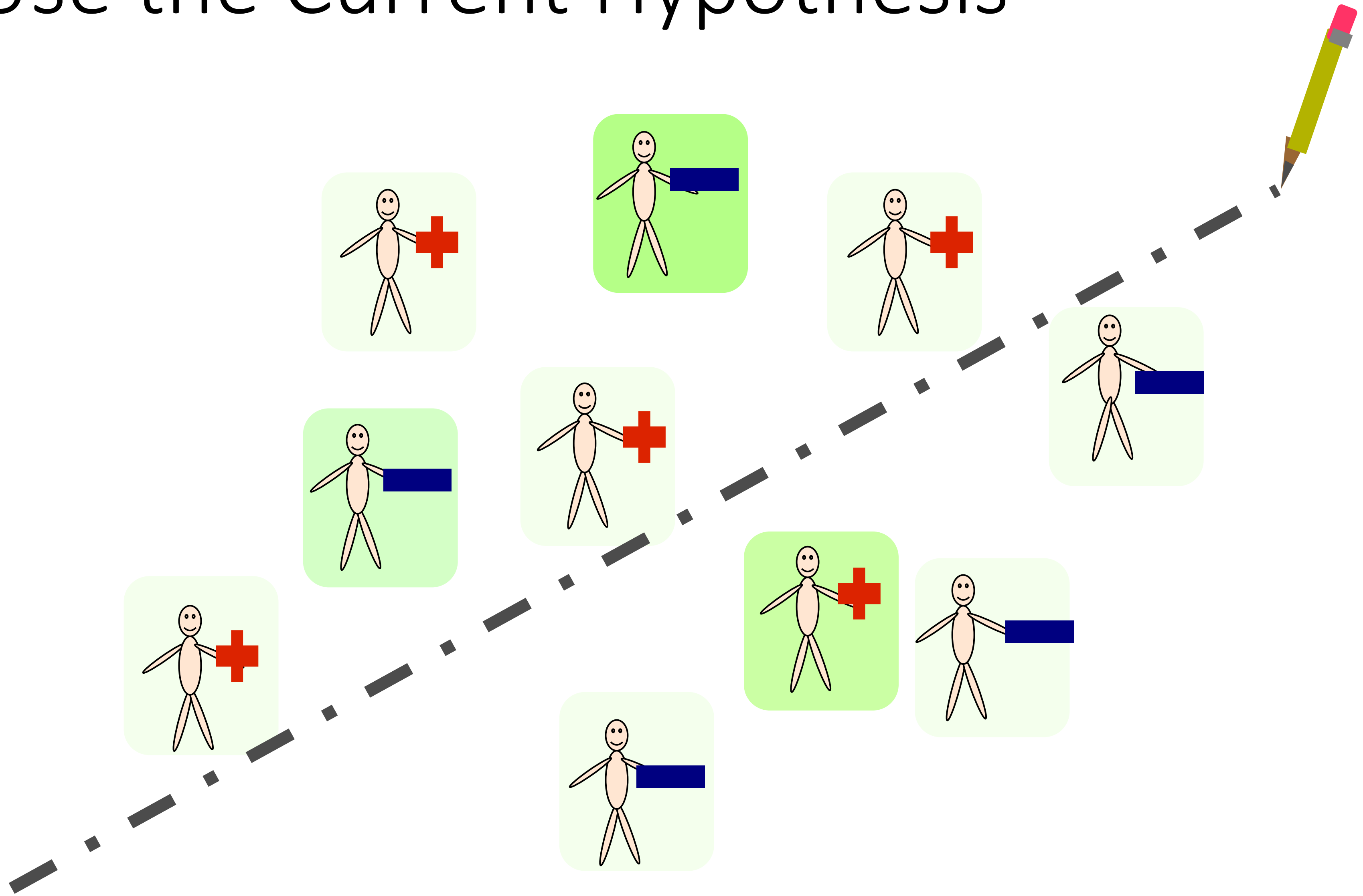
Challenge 1: How to deal with **biases**?

Challenge 2: Which data is more **useful**?

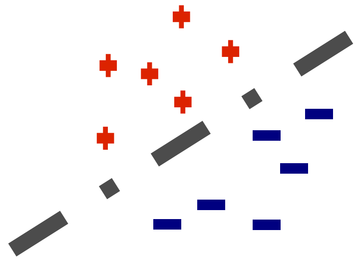
How to Assess Value/Price of Data?



Use the Current Hypothesis



Intuitive Example



- Perceptron algorithm [Rosenblatt, 1958]
  - An online algorithm for learning the linear classifier
  - For each arriving point:
    - If the current hypothesis is right, do nothing
    - If the current hypothesis is wrong, update the hypothesis
  - If there exists a perfect hypothesis
    - The algorithm makes at most ***1/(margin)²*** mistakes

- Pay for mistakes!

Main Result

- For a general class of learning algorithms (FTRL, e.g., online gradient descent, and multiplicative weight updates), our mechanism achieve

measure of “problem difficulty”, in [0,1].

measure of how “good” our mechanism is

$$E\ \mathrm{loss}\left(\bar{h}\right) \le E\ \mathrm{loss}\left(h^*\right) + O\left(\sqrt{\frac{\gamma}{B}}\right)$$

our hypothesis

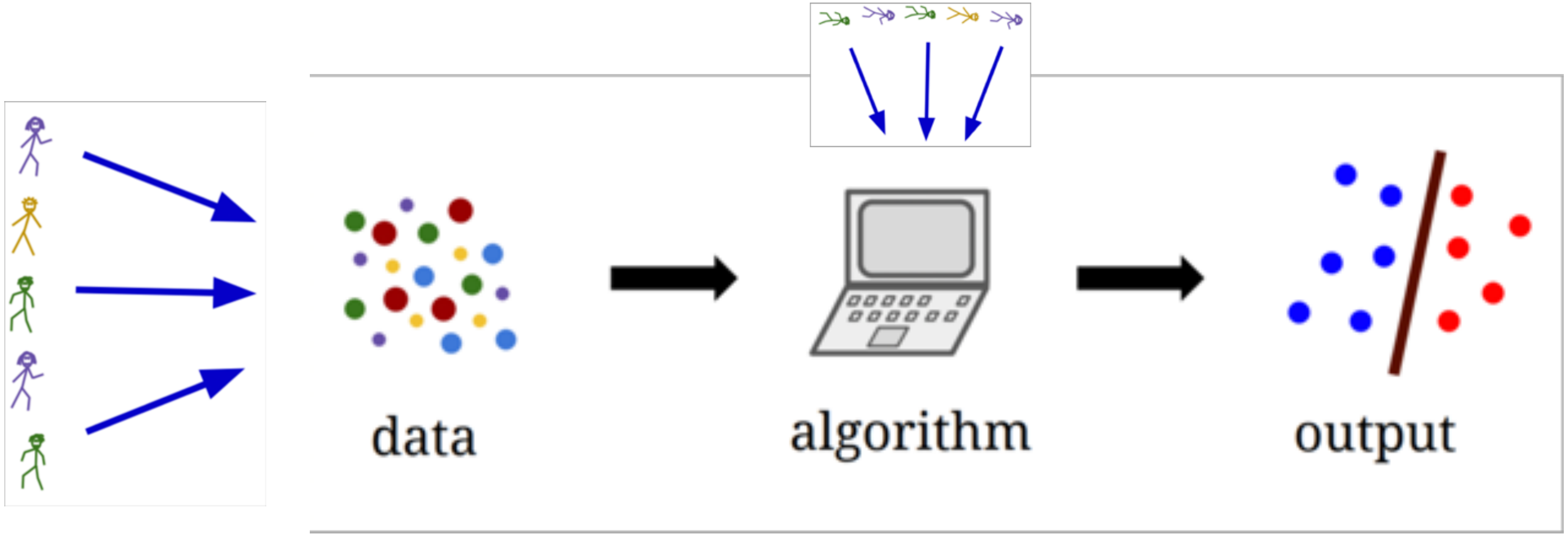
optimal hypothesis

Budget

- For any mechanism,

$$E\ \mathrm{loss}\left(\bar{h}\right) \ge E\ \mathrm{loss}\left(h^*\right) + \Omega\left(\frac{\gamma}{\sqrt{B}}\right)$$

Learning with Humans in the Loop

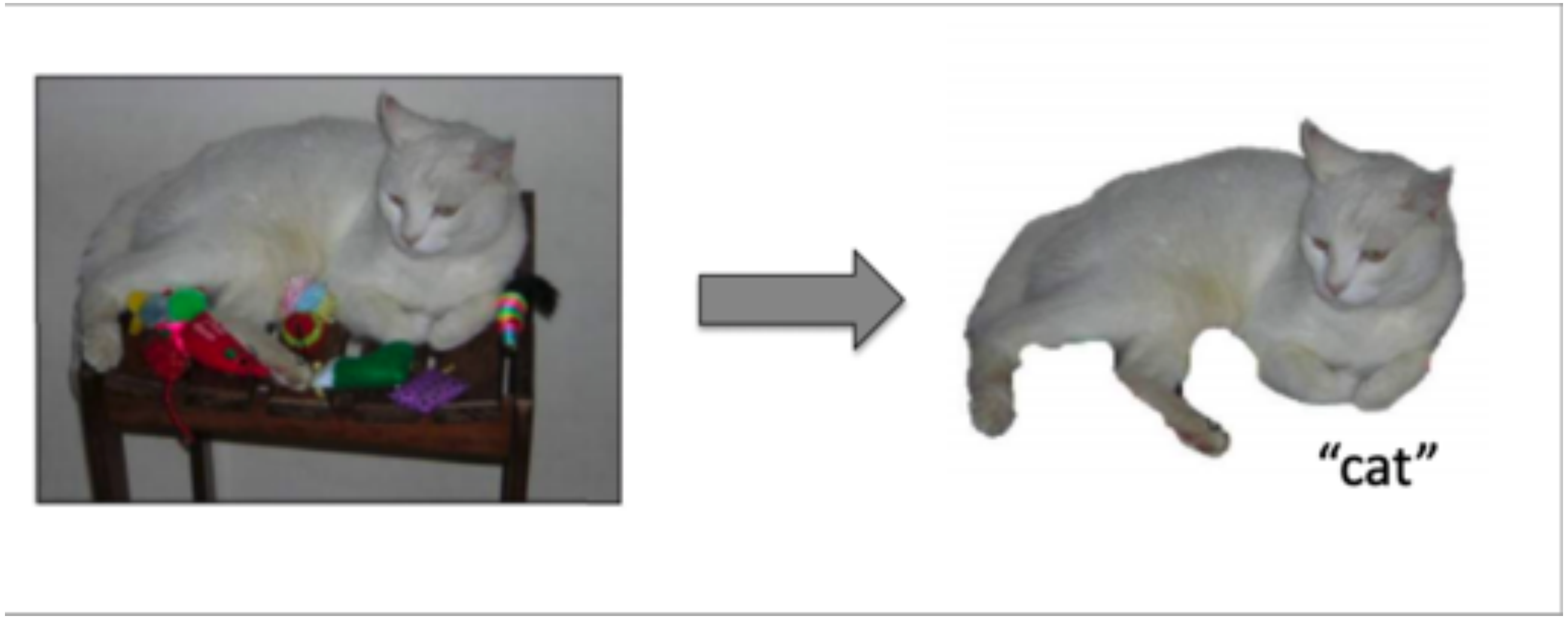


We have focused on the aspect of “data generated by humans”
Make strong i.i.d. assumptions on the data generation process

2. Can humans be involved in the algorithm as well?

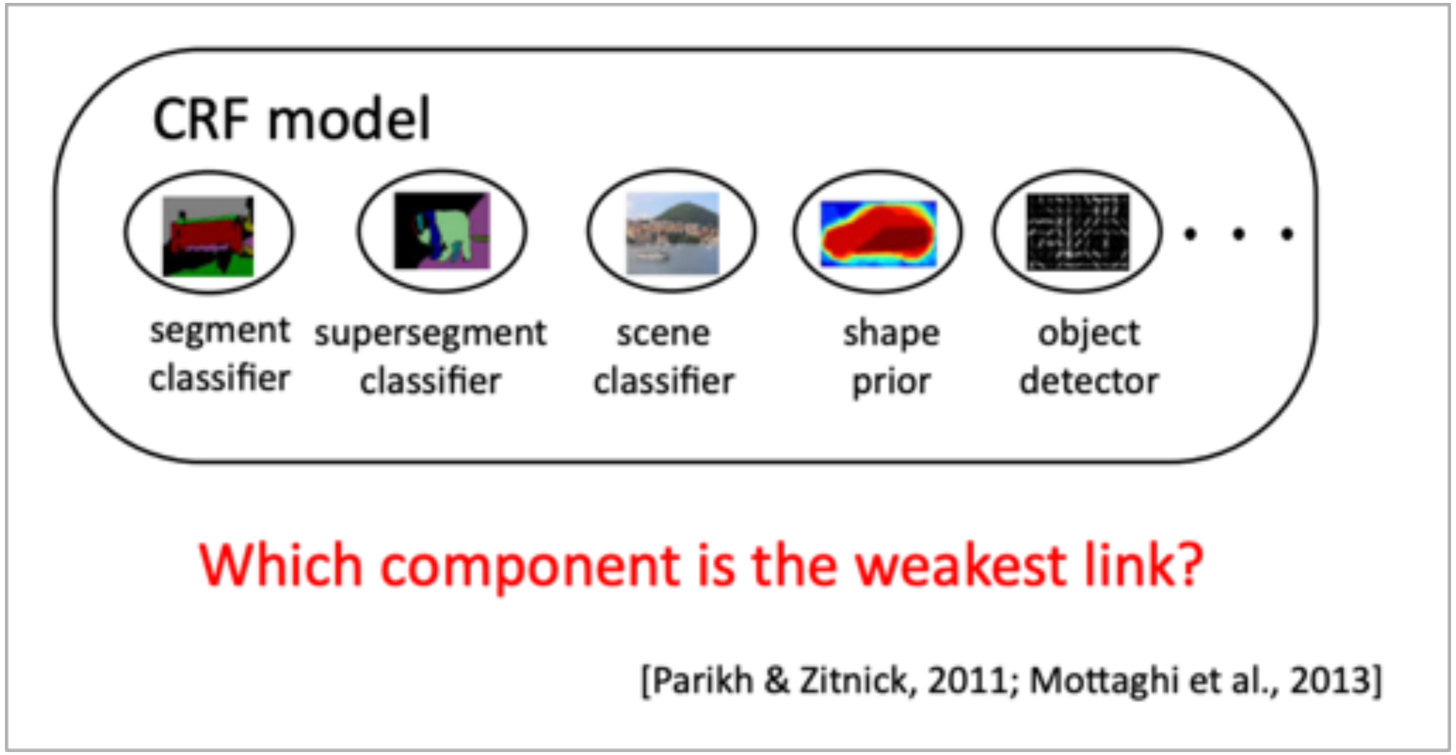
Human Debugging

- Semantic segmentation: partition an image into semantically meaningful parts, and label each part.

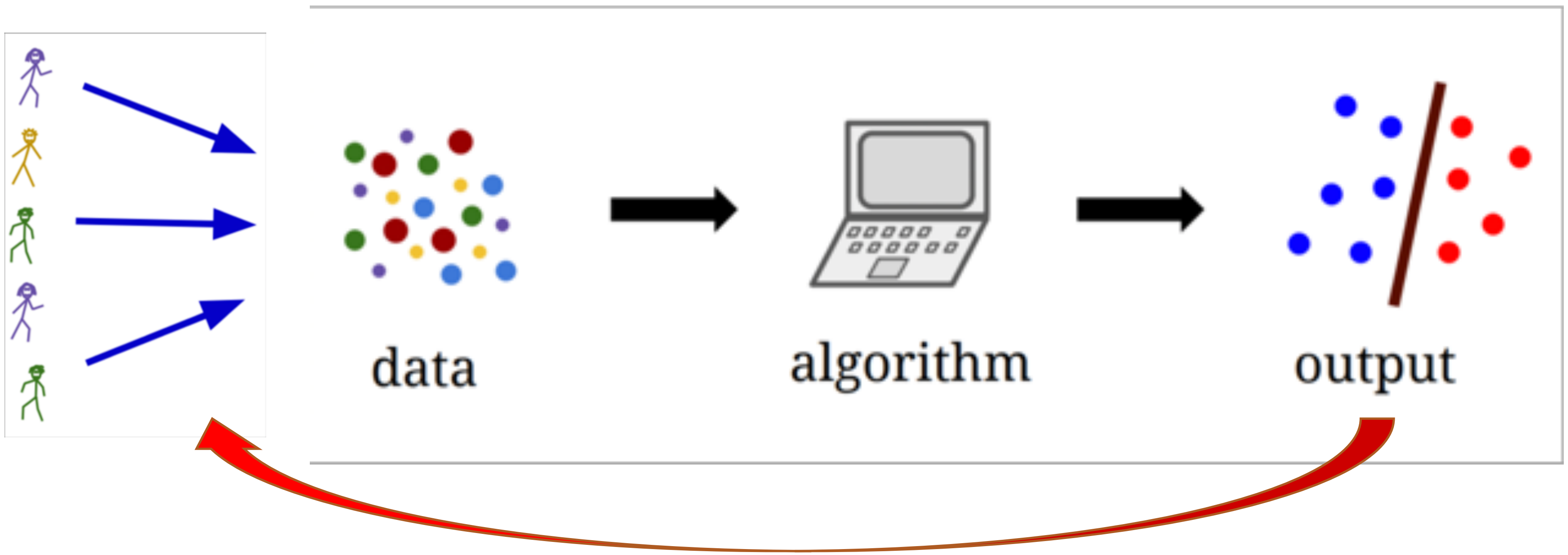


Human Debugging

- Semantic segmentation: partition an image into semantically meaningful parts, and label each part.



Learning with Humans in the Loop



We have focused on the aspect of "data generated by humans"
Make strong i.i.d. assumptions on the data generation process

3. What if the output has impacts to humans?

Counterfactual Machine Learning

- Consider you are a search engine
  - You use the click data to estimate how relevant a website is to a search
  - Determine how to rank them based on the estimation

- What’s wrong here?
  - The data you collected depends on how you rank
    - Websites ranked higher gets more clicks
  - Need to infer “what if” we use a different policy

- Standard approaches: A/B-Testing => Too many possibilities
- Counterfactual ML => Develop user models to answer the “what if” question

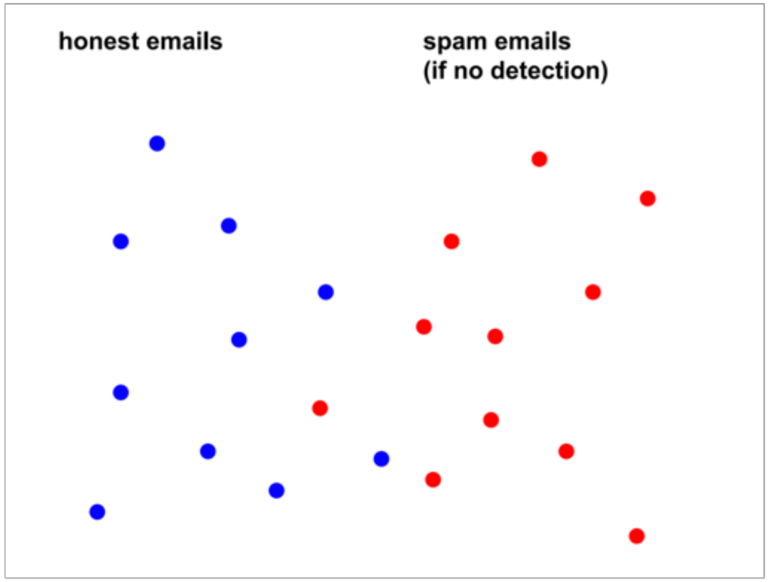
Strategic Machine Learning

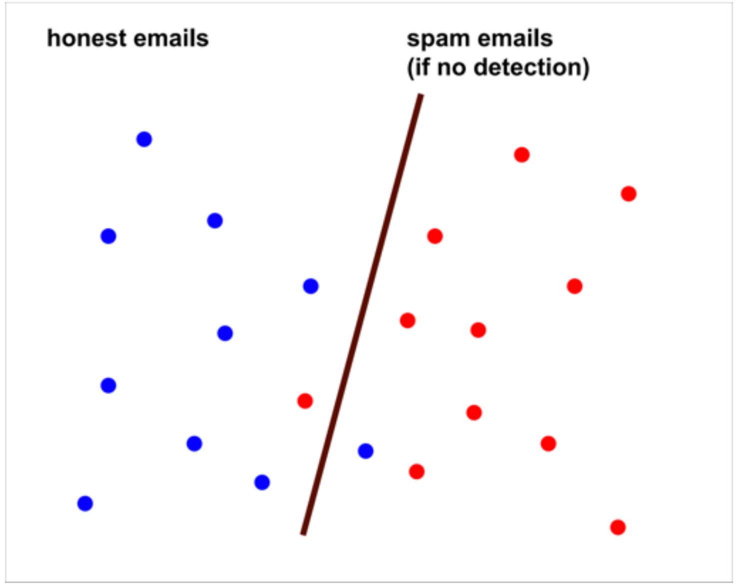
Goodhart’s law:

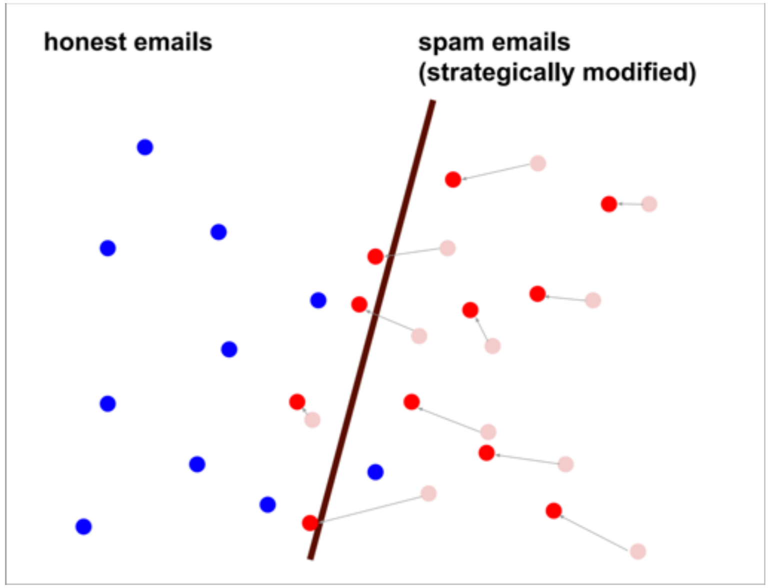
> “If a measure becomes the public’s goal,
> it is no longer a good measure.”

Strategic Machine Learning

- Spam classification example as in last lecture



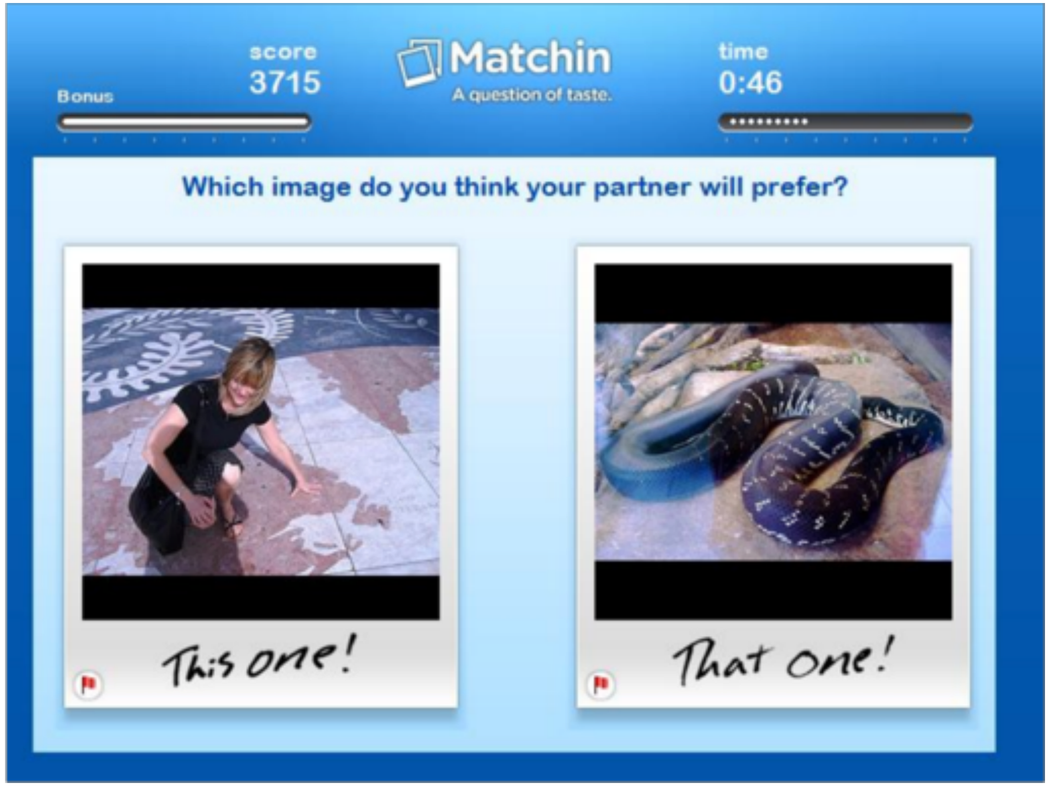




- More examples: School admission, Job offer determination, etc

Ethical Issues

- A Game for Collecting User Preferences on Images



Matchin: Eliciting User Preferences with an Online Game. Hacker and von Ahn. CHI 2009.

Ethical Issues

- Which one do you like

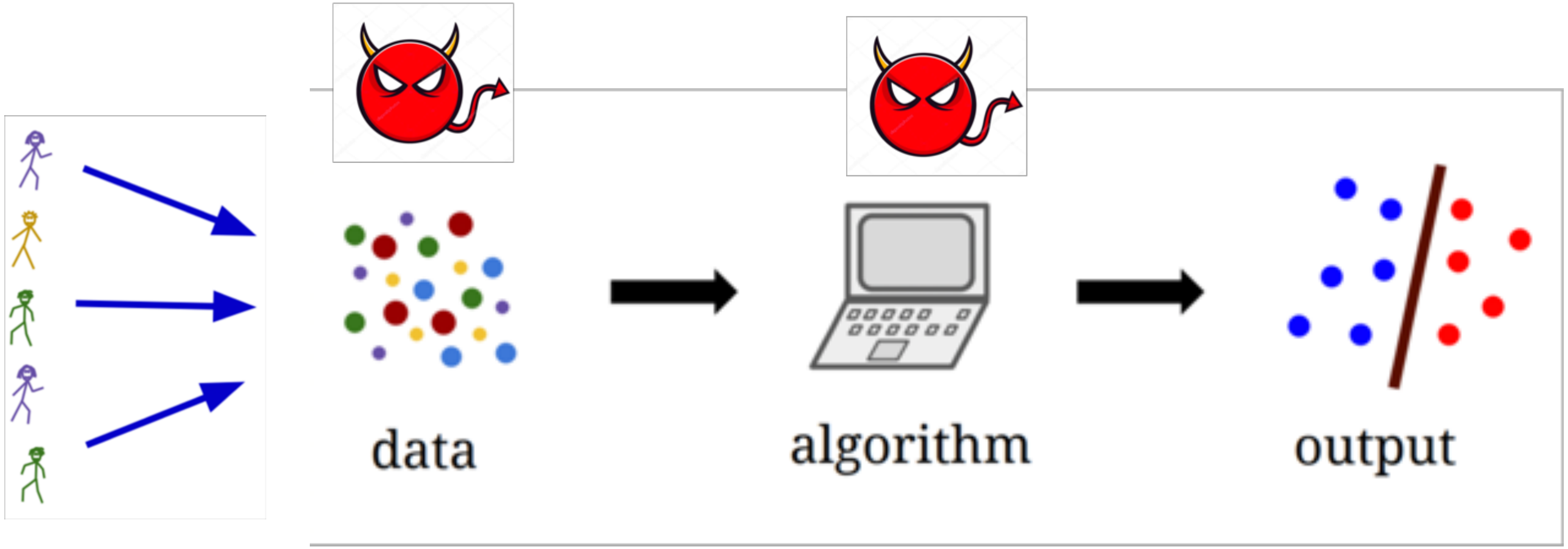
Matchin: Eliciting User Preferences with an Online Game. Hacker and von Ahn. CHI 2009.

Ethical Issues

- Building Gender Models using Images
- Ask MTurk workers to compare 10 pairs of pre-selected images.
  - Accuracy for guessing the gender correctly: 78.3%
- Workers are not only contributing data. They might be sacrificing some privacy as well.

Matchin: Eliciting User Preferences with an Online Game. Hacker and von Ahn. CHI 2009.

Learning with Humans in the Loop



We have focused on the aspect of “data generated by humans”
Make strong i.i.d. assumptions on the data generation process

4. What if some users try to sabotage the system.

Adversarial Machine Learning (CSE544T, SP 2019)

(a) (b) (c)

Can an adversary "inject" a small amount of data to break ML algorithms?