Machine-Learning Aided Peer Prediction

Yang Liu and Yiling Chen

yangl, yiling@seas.harvard.edu

Harvard University

June 28th, 2017

HARVARD
JOHN A. PAULSON
SCHOOL OF ENGINEERING
AND APPLIED SCIENCES

Information Elicitation without Verification

CNN politics

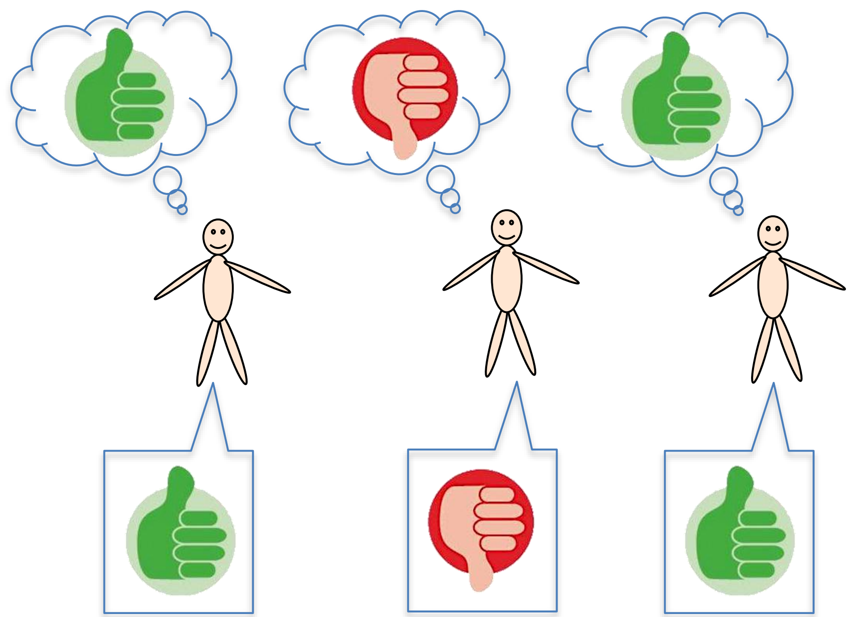
Trump's outlook

Adult content free?

Yes

No

Private opinions



Goal: collect truthful reports

Challenge and solution

Costly or impossible to verify reports against observable ground truth

Theoretical solution: peer prediction

- Under certain assumptions, a family of mechanisms can truthfully elicit private signals at an equilibrium. [Prelec 04, Miller et al. 05, Jurca & Faltings 09, Witkowski & Parkes 12, Radanovic & Faltings 13, Witkowski 13, Dasgupta & Ghosh 13, Zhang & Chen 14, Shnayder et al. 16, Kong & Schoenebeck 16]

Basic Idea of Peer prediction

Verifies the reports against one another

Rewards = how well each report predicts other reports

Your report	Other random report	Your payment
		1.50
		0.10
		0.30
		1.20

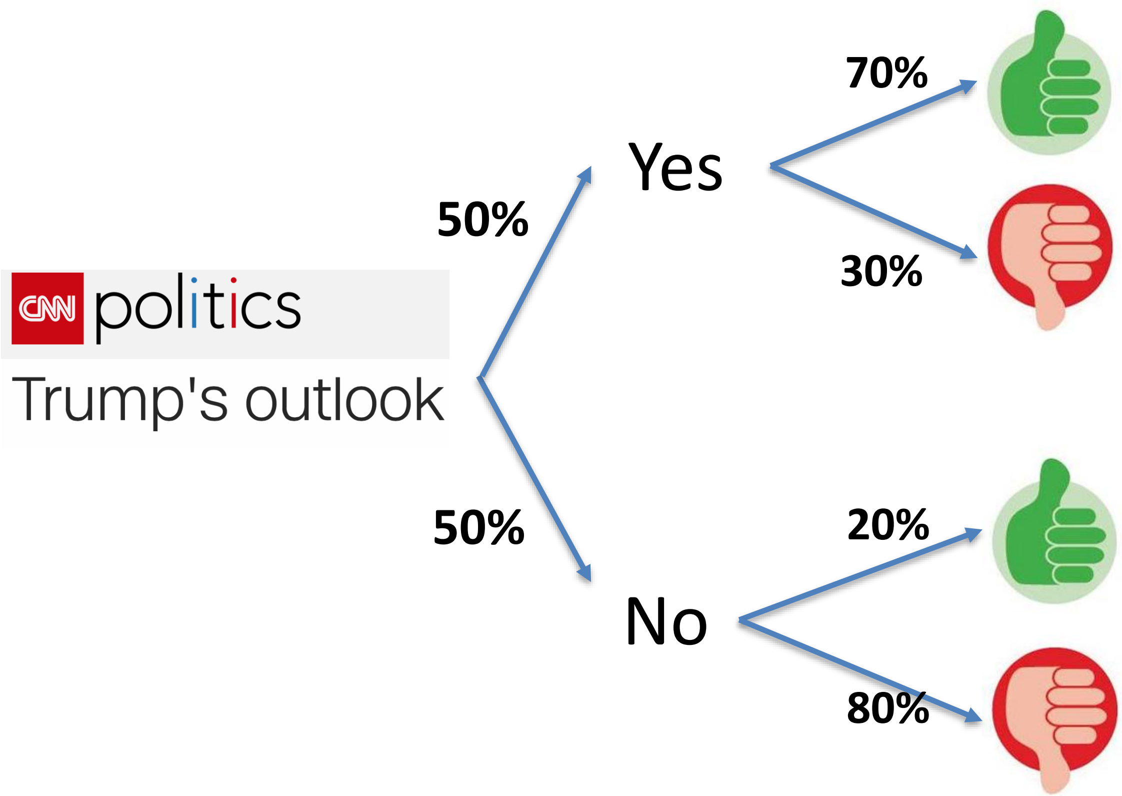
Truthful Equilibrium (under certain assumptions)

Drawbacks of Peer Prediction

- Redundant assignments (inefficiency?)
- Assump. on prior knowledge of observation model
- Known to have uninformative equilibria

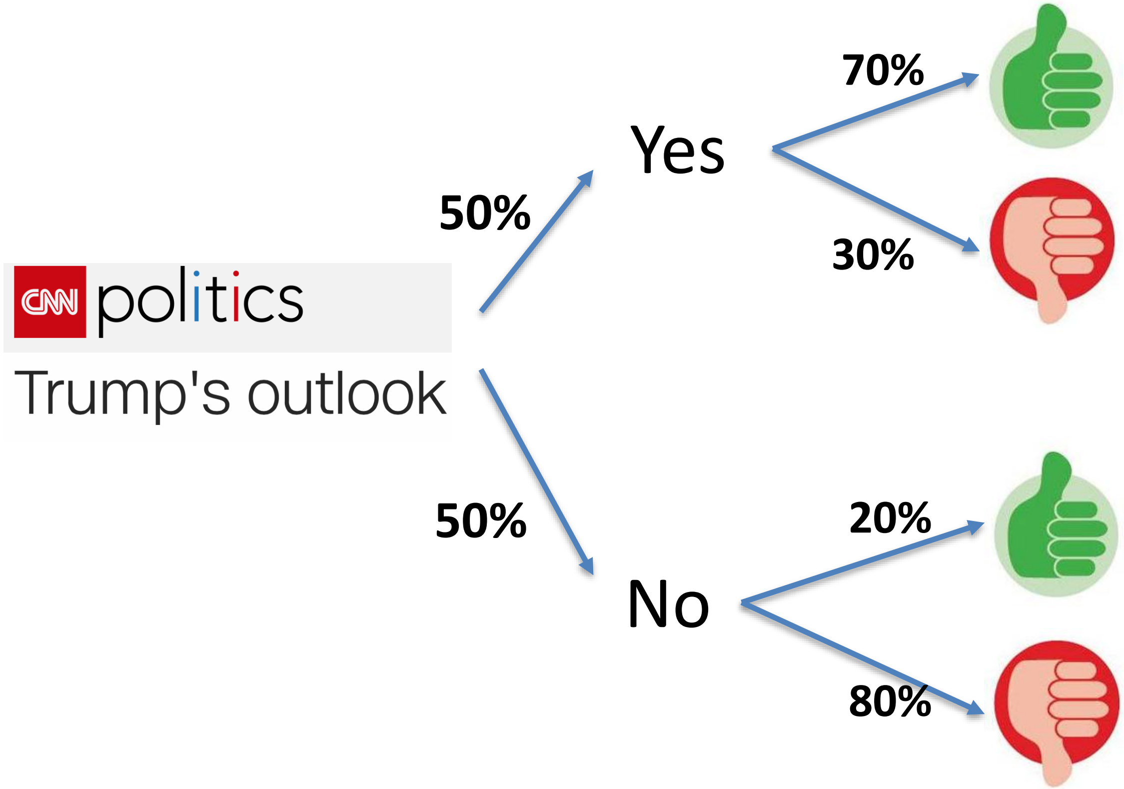
Your report	Other random report	Your payment
		1.50
		0.10
		0.30
		1.20

Ground-truth and opinion



- Common prior
- Common knowledge for all participants and designer

What happened in peer prediction?



Reports

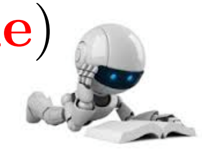
x	y	$\tilde{y}$	$\hat{y}$
feature	***label***	***Human label***	***Collected label***

What happened in peer prediction

- PAYMENT (an agent’s report, a peer agent’s report)
- What we really need is a “prediction” on y (Machine learning)
- Feature vector x is largely ignored.How to associate x with y? (Machine learning)

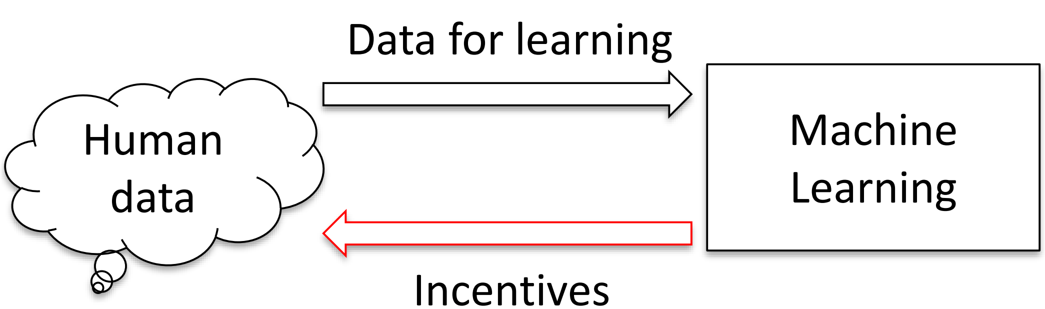
- **A Machine-Learning aided approach**
  - Leveraging the correlation between x and y.

$$\mathrm{PAYMENT}(\hat{y}_i, \mathbf{Machine})$$



Advantages & achievements

- Budget efficient: no need to reassign every task at least twice; leverages more information on x.
- Remove requirement of knowing the observation model.
- Strictly truthful BNE.
- Uninformative strategies no longer form an equilibrium.
- Induces effort exertion.



Model

- N data points to be assigned to label x_i, $(x_i, y_i) \in \mathcal{D}$.
- Binary class: $y_i \in \{+1, -1\}$. The MD knows the priors

$$\Pr(y_i = +1), \Pr(y_i = -1)$$

- Workers' observations follow flipping error model

$$e_{+1} := \Pr(\tilde{y}_i = -1 | y_i = +1), \; e_{-1} := \Pr(\tilde{y}_i = +1 | y_i = -1)$$

  - Homogeneous workers (parts of the results generalize to heterogeneous workers too)
  - Bayesian informative workers ⇔ $e_{+1} + e_{-1} < 1$
- Each data is assigned to exactly one worker, except a small fraction of them which are assigned to two.

Challenges

- Q1: How to learn a good predictor from human generated data?
- Q2: How to design scoring function to incorporate the machine prediction?
- Q3: How to learn the noise rates in agents' observations?

Q1: Learning with noisy data

- Suppose agents truthfully report, with the noisy $\{\mathbf{x}_j, \tilde{y}_j\}$, how to train a predictor ?
- Direct risk minimization will incorporate the noise

$$\tilde{f}^* = \operatorname{argmin}_f \xcancel{\hat{R}_l(f) := \frac{1}{N-1}\sum_{j \neq i}} l(f(\mathbf{x}_j), \tilde{y}_j).$$

- Idea: define surrogate loss to remove bias

$$\varphi(t, \tilde{y}) := \frac{(1 - e_{-\tilde{y}})l(t, \tilde{y}) - e_{\tilde{y}} l(t, -\tilde{y})}{1 - e_{+1} - e_{-1}}, \; e_{+1} + e_{-1} < 1.$$

Q1: Learning with noisy data

- Intuitively

$$\mathbb{E}_{\tilde{y}}[\varphi(t,\tilde{y})] = l(t,y), \forall t.$$

- Empirical risk minimization over surrogate

$$\tilde{f}^*_{\tilde{l}} = \mathrm{argmin}_f \hat{R}_{\tilde{l}}(f) := \frac{1}{N-1} \sum_{j \neq i} \varphi(f(\mathbf{x}_j), \tilde{y}_j).$$

- Enough samples => a good prediction $\hat{f}^*(x_i)$
  - Note we need to know the error rates.

Q2: Design of the scoring function

Surrogate loss function serves as a scoring function.

- Score/pay each worker $-\varphi(\hat{f}^*(x_i), \hat{y}_i)$

Reference report as the classifier prediction

$$S(\hat{y}_i, \hat{y}_j) \qquad\qquad \varphi(\tilde{Y}, f(X))$$

Max reward = Min loss

Elicited report as the noisy label

- A strictly truthful BNE
  - **Intuition**: Remove bias in agent's report, so will positively correlate with the classifier's prediction.
- Not the only way to do so : **M**achine **O**utput **A**greement

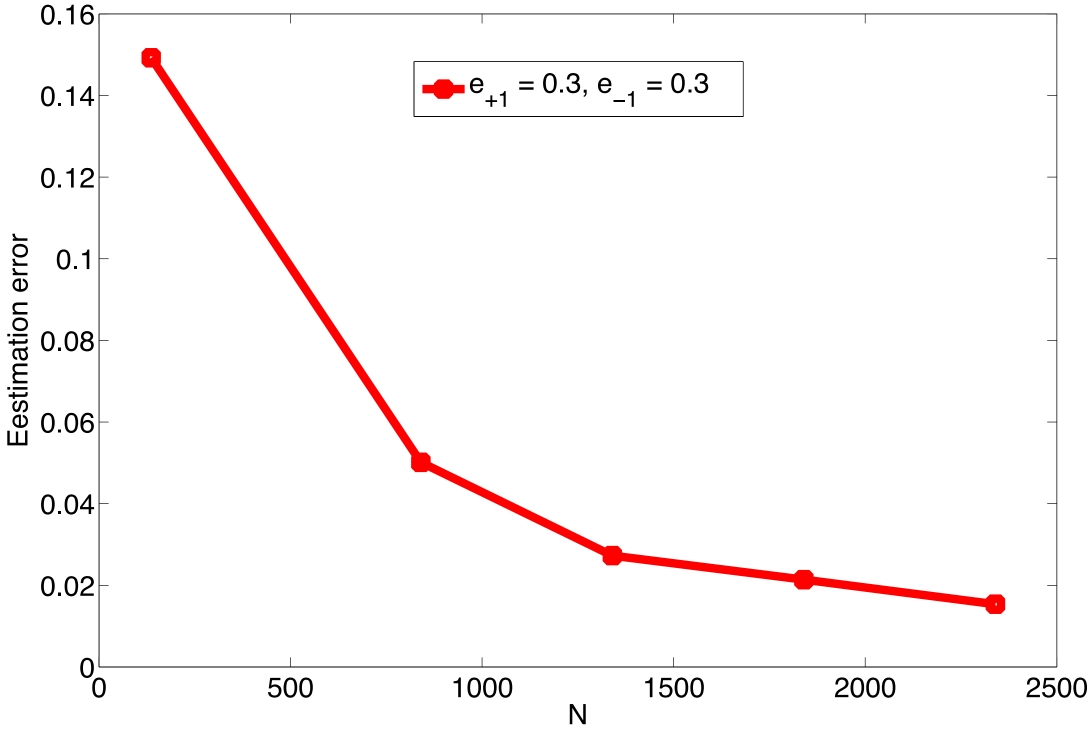
Q3: Learning the error rates

$$\mathcal{P}_+[e_{i,+1}^2 + (1-e_{i,+1})^2] + \mathcal{P}_-[e_{i,-1}^2 + (1-e_{i,-1})^2] = \textbf{Pr(matching)}$$

$$\mathcal{P}_+ e_{i,+1} + \mathcal{P}_-(1-e_{i,-1}) = \textbf{Fraction of -1 labels observed}$$

- **Lemma**: can accurately learn the error rates when agents adopt symmetric pure strategies
  - Upper bound on the number of samples needed.

Q3: Learning the error rates



When the error rates are small enough, the truthfulness will retain.

ML elicitability

We call a data distribution being **ML elicitable** if there exists a mechanism that can learn a classifier and scoring function that has a (strictly)truthful reporting equilibrium.

Theorem: A data distribution is ML elicitable, if

- There exists a concept class with bounded VC dimension
- The optimal classifier performs better than random guess

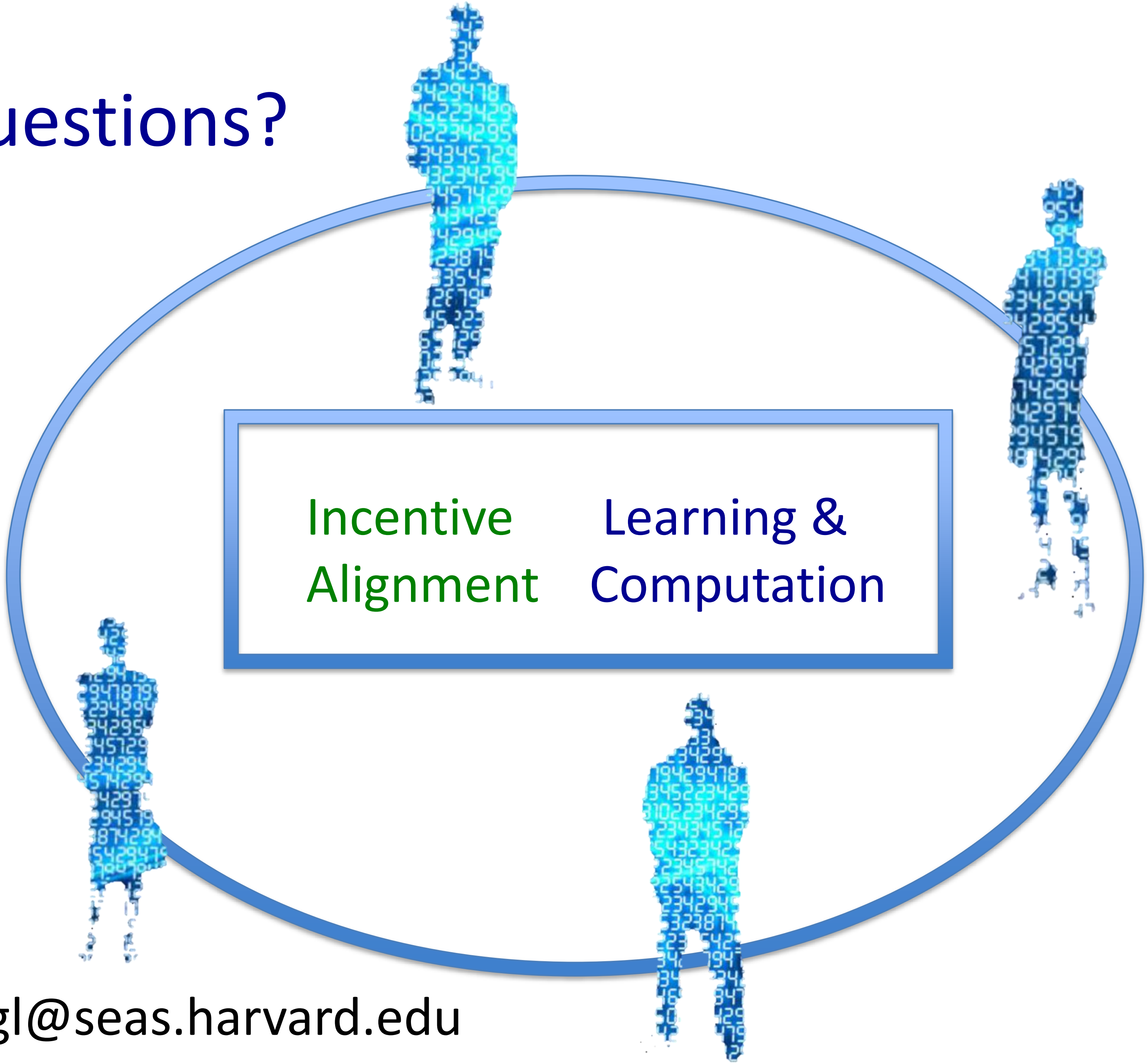
$$\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[1(f^*(\mathbf{x}) \neq y)] < 0.5.$$

- No worse than random guess for each class

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}|y}[1(f^*(\mathbf{x}) \neq y)] \leq 0.5,\ y \in \{-1,+1\}$$

Also extends to effort elicitation case

Questions?



yangl@seas.harvard.edu