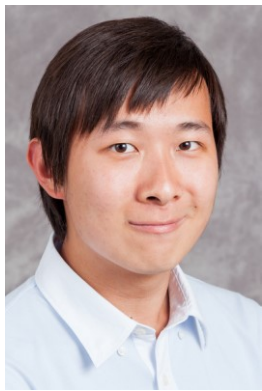# Automated Mechanism Design for Strategic Classification
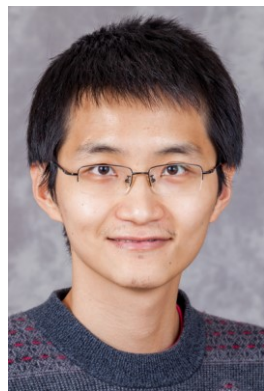
Vincent Conitzer, joint work with:

Hanrui Zhang (Duke)
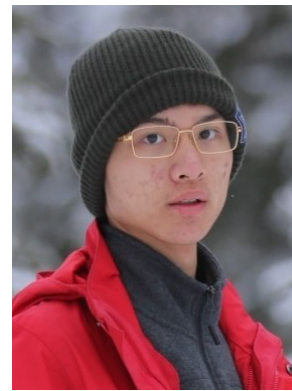
Andrew Kephart (Duke → Instacart)

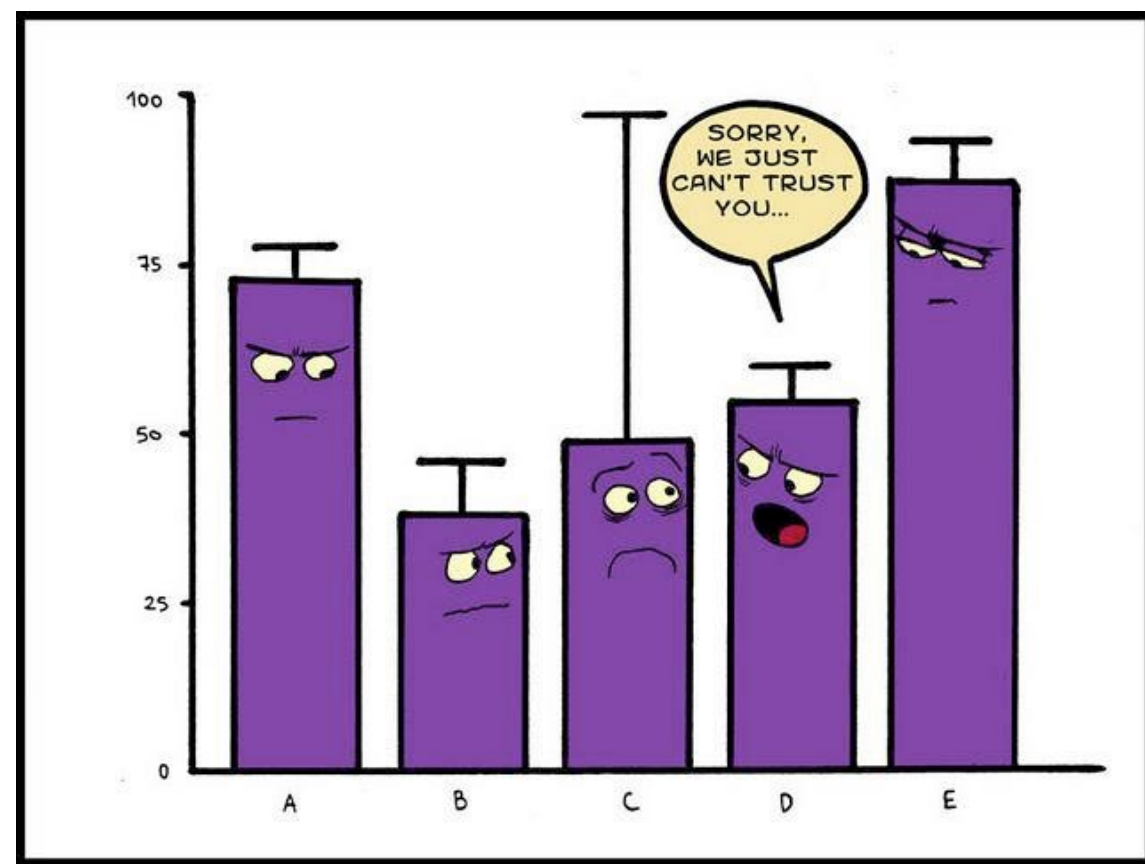Yu Cheng (Duke → UIC)

Anilesh K. Krishnaswamy (Duke)

Haoming Li (Duke → USC)

David Rein (Duke)

AI / algorithms are making decisions about us!

- Will you get **a loan**?
- Will you get **a job**?
- Will you get **a date**?
- Will you get **out on bail**?

China's Tech Giants Charge Into Financial Services

| Shareholding ratio | Meituan Dianping Food-delivery | Didi | Xiaomi Smartphone-maker | JD.com E-commerce |

Get started   Open in app

AI in Dating Apps: The Changing Face of Online Dating Industry

≡ WIRED   BACKCHANNEL   BUSINESS   CULTURE   GEAR   IDEAS   SCIENCE   SECURITY   SIGN IN   SUBSCRIBE

TOM SIMONITE   BUSINESS   02.19.2020 08:00 AM

In depth: Want a loan? China's
Algorithms Were Supposed to Fix the Bail System. They Haven't
tech giants are at your service
A nonprofit group encouraged states to use mathematical formulas to try to eliminate racial inequities. Now it says the tools have no place in criminal justice.

Huge troves of user data prove
invaluable for insurance,
healthcare and other services

https://asia.nikkei.com/Spotlight/Caixin/In-depth-Want-a-loan-China-s-tech-giants-are-at-your-service

PHOTOGRAPH: GUY CALI/GETTY IMAGES

| Fund sales | Hegeng Chuancheng Fund Sales 90% | Du Xiaoman Financial |

"How AI and big data helped China's tech giants dominate consumer finance" [South China Morning Post, 11-26-2020]

In Ant's case, the terms of the loan will be largely determined by Ant's Zhima credit, a credit-scoring system based on a user's digital footprint, including records from payment systems and even whether he or she returned a shared power bank on time. If a consumer is willing to offer more personal information, such as their record of house purchases or even details of their professional LinkedIn profile, he or she can potentially get a higher score at Zhima Credit.

[…]

"Birds of a feather flock together. Similar people usually have the same kind of risk – those correlations could include whether they visit similar apps and websites, or receive similar calls," he said.

And tech companies currently gather more data on their users than almost any other industry – handing them a natural advantage.

"Artificial Intelligence in Payments: 1-second AI loan decisions" [PaymentGenes, 18-02-2020]

How **Alibaba** and **Tencent** feed data into loans?

Alibaba
Microbusinesses, farmers, and others using e-commerce

Tencent
Individuals

Records of smartphone-based payments
(transaction amounts, utility bills, etc.)

User IDs, registered profiles, and owned assets

Assessment of e-commerce transactions
Assessment of purpose of funds and references
(in case of farmers)

Record of calls and texts to confirm personal relations and guard against organized crime

AI decides on loan amounts, periods, and other terms

PaymentGenes

"Boost Your Chances of Getting Your Personal Loan Approved" [NerdWallet, June 17, 2020]

2. Rebalance your debts and income

[…]

Consider selling liquid assets such as stocks held in taxable accounts. Using the proceeds toward high-interest consumer debts should get you a higher rate of return, says Alison Norris, advice strategist and certified financial planner at personal finance company SoFi.

Boosting your income and lowering your debt improves your debt-to-income ratio, which is the percentage of your monthly debt payments divided by monthly income. Not all lenders have strict DTI requirements, but a lower ratio shows that your current debt is under control and you can take on more.

3. Don't ask for too much cash

Requesting more money than what you need to reach your financial goal can be seen as risky by lenders, says Norris.

"Look at the reason why you're asking for the loan, tie a specific dollar amount to that financial need, and only ask for that amount," she says.

[…]

"What to Put on a Resume (and What Not to Put on a Resume)"
[Jobscan, July 23, 2020]

5. Irrelevant Experience and Skills

A great resume is customized for the job to which you're applying. However, removing years of irrelevant experience can cause work history gaps. And so the key becomes finding ways to connect your experience to the job. It's best if you can do this with relevant hard skills and keywords. Here are some blog posts that can help:

- How to Write a Career Change Resume
- How to Write a Resume with No Experience

8. Untruths

A recent study found that 78% of applicants admit to lying on their resumes and that 66% of hiring managers didn't actually care. That may be some comfort if you currently have, ahem, untruths on your resume but we suggest taking a hard look and asking yourself if you're willing to get caught in a lie. Many companies will vet new hires and a lie could cost you the job.

* 0 details about survey…

"AI May Be Reviewing Your Resume. Here's How to Stop It From Rejecting You"

Inc.

watch out for titles that only exist at a specific company because AI may not recognize them.

"Some people are storytellers, but that's not necessarily good for AI because AI is designed to filter this out." Instead, focus on relevant facts, he advises. "For example: 'I led a team of 20 people for three years and we increased the company's revenue by 40 percent.' This is something AI can digest more easily."

AI might misread your resume and think you haven't been using your skill or skills lately because the skill itself is only named once, he warns. In this case, repetition is your friend. "We need to make sure it repeats across different experiences and positions," Mordechay says.

"How machines are changing the way companies talk" [VentureBeat, December 27, 2020]

The analysis found that companies are adapting their language in forecasts, SEC regulatory filings, and earnings calls due to the proliferation of AI used to analyze and derive signals from the words they use. In other words: Businesses are beginning to change the way they talk because they know machines are listening.

"More and more companies realize that the target audience of their mandatory and voluntary disclosures no longer consists of just human analysts and investors. A substantial amount of buying and selling of shares [is] triggered by recommendations made by robots and algorithms which process information with machine learning tools and natural language processing kits," the paper reads. "Anecdotal evidence suggests that executives have become aware that their speech patterns and emotions, evaluated by human or software, impact their assessment by investors and analysts."

Some takeaways

- Some actions change the underlying state of the world (not the focus here)

- Some amount of presenting the information differently might be desirable

- There may be incentives to lie…

- … but some lies would be caught

Classifying strategic agents

[Kephart & C. AAMAS 2015; Hardt, Megiddo, Papadimitriou, Wootters ITCS 2016; …]

Data from agents is used to train classifier…

data from agents to be classified

-$2000 in account
A is my friend

Don't lend if debt exceeds $1000
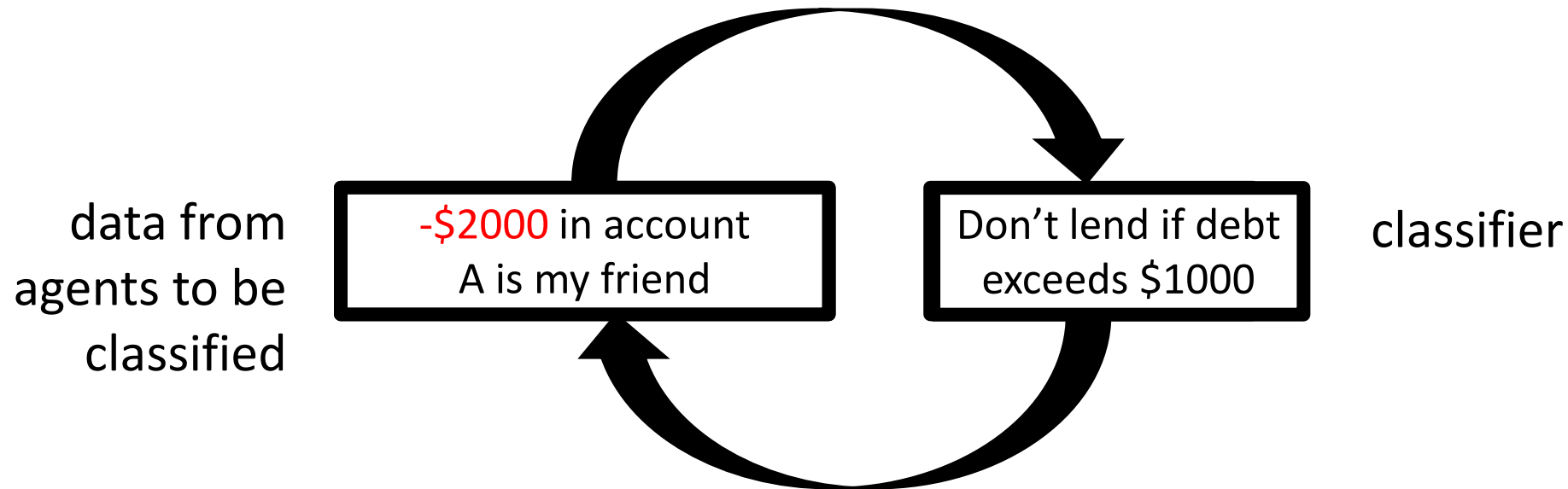
classifier

… but agents best-respond to the classifier in submitting data

setting is not just **adversarial** (zero-sum)

Why should we even get involved in this?

- "These processes are biased and/or otherwise awful."
- "A human should be in the loop."
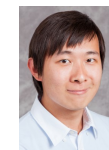- "Give an underdog a chance!"
- "AI outsmarting human beings is scary."

After an audit of the algorithm, the resume screening company found that the algorithm found two factors to be most indicative of job performance: their name was Jared, and whether they played high school lacrosse.
https://qz.com/1427621/companies-are-on-the-hook-if-their-hiring-algorithms-are-biased/

Former Boston Public Schools superintendent, Thomas Payzant, justified the change to a strategy-proof Deferred Acceptance student-school allocation mechanism by explaining, "a strategy-proof algorithm 'levels the playing field' by diminishing the harm done to parents who do not strategize or do not strategize well."
https://blogs.cornell.edu/info4220/2016/03/17/the-importance-of-strategy-proof-mechanisms-in-the-school-admissions-process/

https://www.cnbc.com/2019/06/26/robots-could-take-over-20-million-jobs-by-2030-study-claims.html

https://giphy.com/gifs/heute-show-yes-celebration-3o75279fBGgbO848ZW

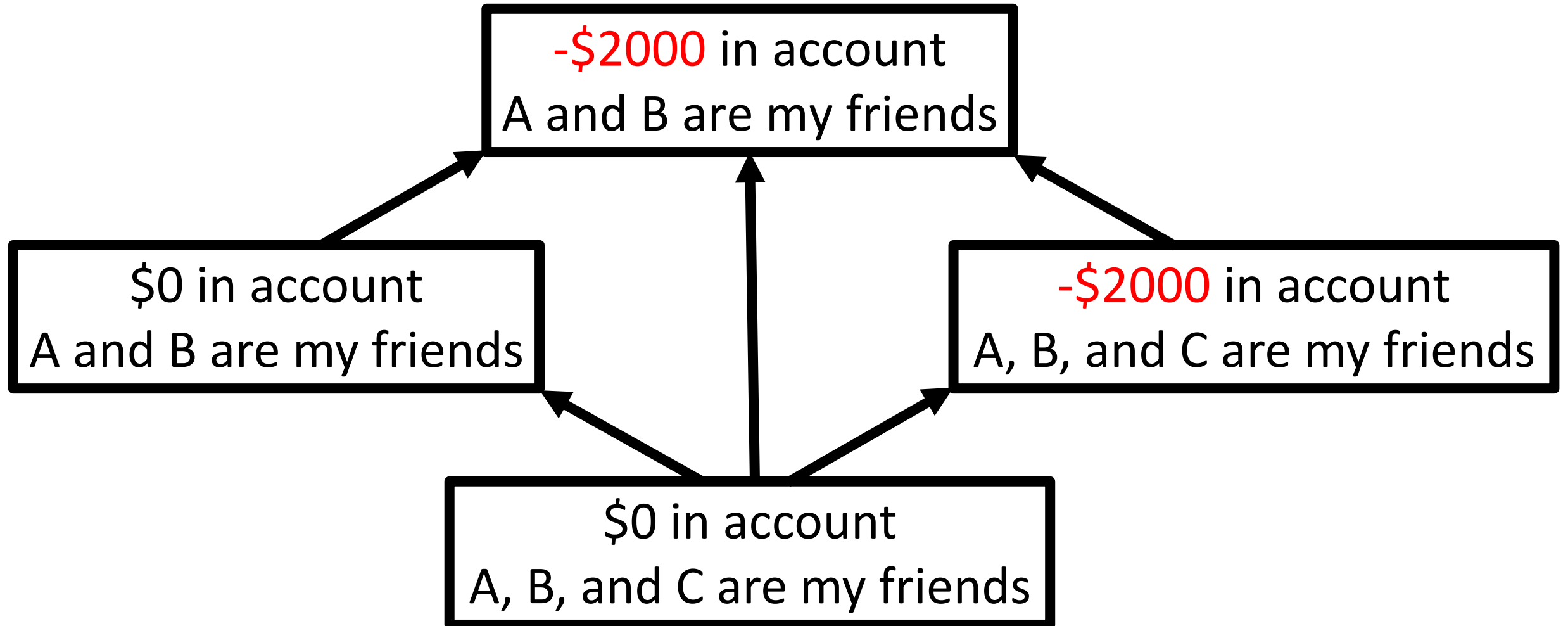Steven Jecmen   Hanrui Zhang   Ryan Liu   Nihar Shah   Fei Fang

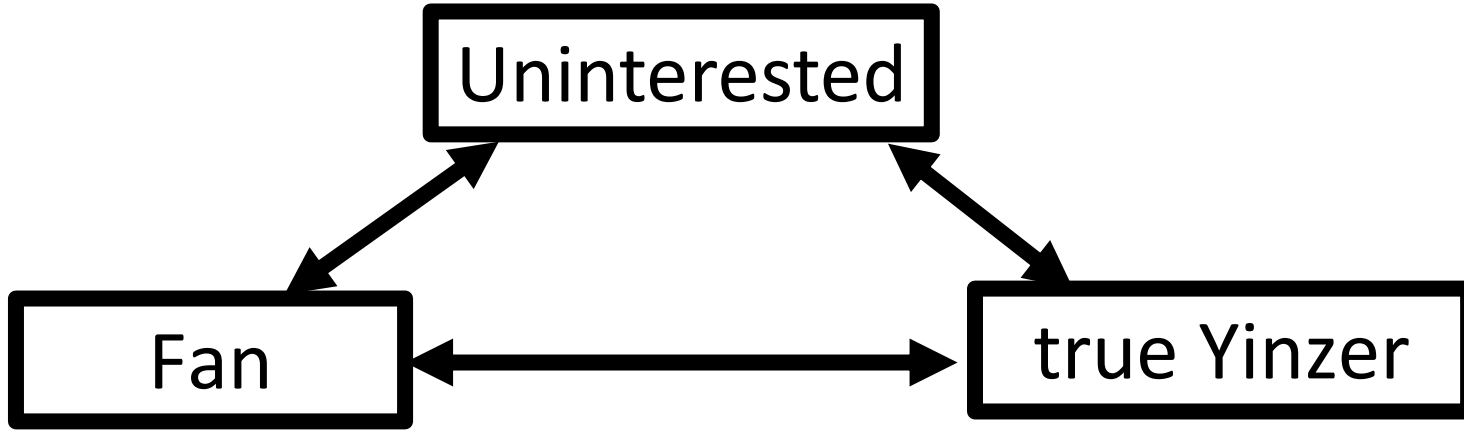# Models of (mis)reporting: direct revelation
Agent's **type** = feature values

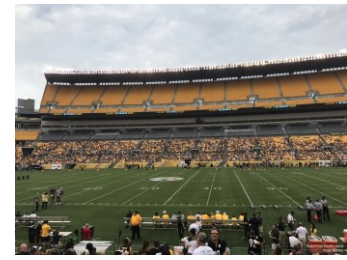# Interlude: Mechanism design for traditional applications

from rateyourseats.com

Great   **Decent**

Selling tickets to a Steelers game

Uninterested

Fan ⟷ true Yinzer

- Three allocations: Great seat, Decent seat, No seat
- $v_U(G)=v_U(D)=v_U(N)=0$
- $v_F(G)=200$, $v_F(D)=100$, $v_F(N)=0$
- $v_Y(G)=500$, $v_Y(D)=200$, $v_Y(N)=0$

- A mechanism:
- U gets N, pays 0
- F gets D, pays 50
- Y gets G, pays 300

Incentive compatible:
No type benefits from misreporting

Variants

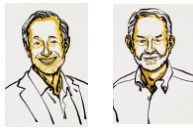|	unlimited misreporting	partial verification / costly misreporting
identical preferences	trivial / can't do much	some classification settings
distinct preferences	**traditional applications**	other classification settings

Nobel Prizes in Economics:
2007 (mechanism design)!
2012 (matching mechanisms)!
2020 (auction mechanisms)!

Hurwicz, Maskin, Myerson   *Roth, Shapley*   *Milgrom, Wilson*

Mingyu Guo
(Duke → U. Liverpool → U. Adelaide)

Angelina Vidali
(Duke → U. Athens)
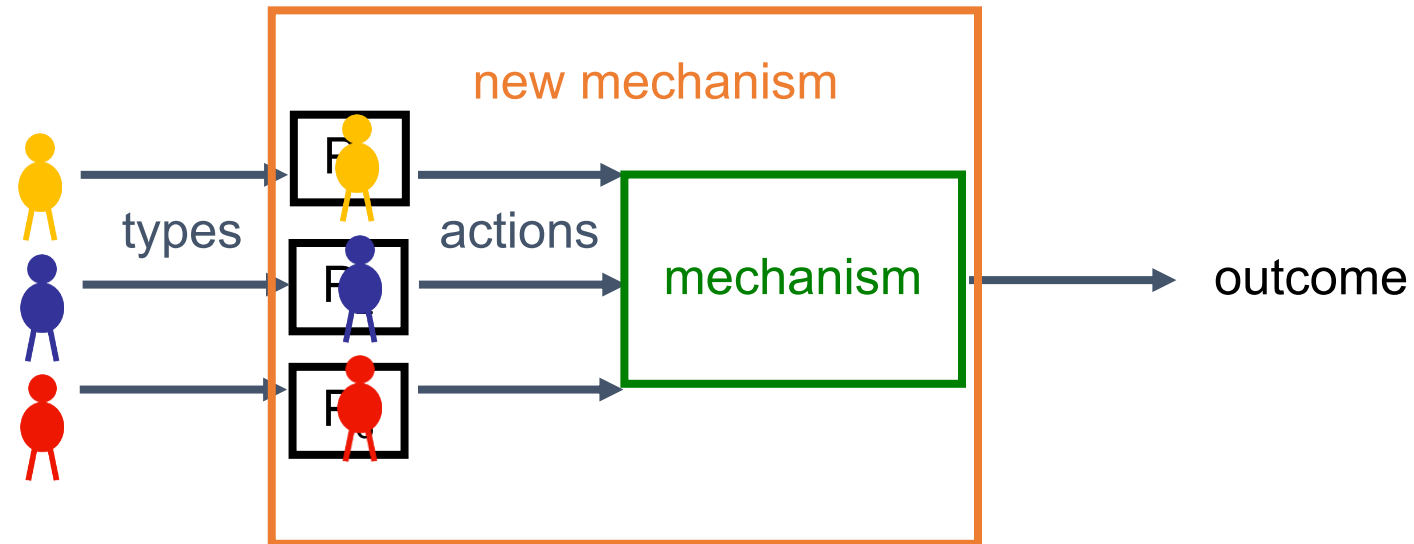
Troels Bjerre Lund (f. Sørensen)
(Duke → ITU Copenhagen)

Melissa Dalis
(Duke → Square → Uber → Mindstrong)

Michael Albert
(Duke → U. Virginia (Darden School of Business))
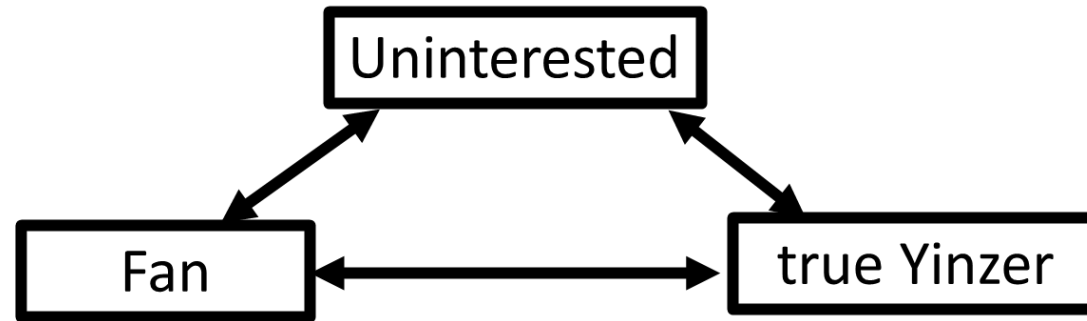
Revelation Principle

- *If* any type can report any (other) type, then it is *without loss of generality* to consider IC mechanisms

Automated mechanism design [C. & Sandholm UAI 2002 and subsequent work] -- example

INPUT

- Three allocations: Great seat, Decent seat, No seat
- $v_U(G) = v_U(D) = v_U(N) = 0$
- $v_F(G) = 200$, $v_F(D) = 100$, $v_F(N) = 0$
- $v_Y(G) = 500$, $v_Y(D) = 200$, $v_Y(N) = 0$

- *Probability distribution:* .3U, .4F, .3Y

- *Other details:* objective (revenue), randomization allowed (yes), …

OUTPUT

- A mechanism:
- U gets N, pays 0
- F gets D, pays 50
- Y gets G, pays 300

Automated mechanism design example continued

Maximizing revenue in Steelers tickets example

```
maximize
0.3pi_1_1 + 0.4pi_2_1 + 0.3pi_3_1
subject to
p_t_1_o1 + p_t_1_o2 + p_t_1_o3 = 1
p_t_2_o1 + p_t_2_o2 + p_t_2_o3 = 1
p_t_3_o1 + p_t_3_o2 + p_t_3_o3 = 1
0p_t_1_o1 + 0p_t_1_o2 + 0p_t_1_o3 - pi_1_1 >= 0
200p_t_2_o1 + 100p_t_2_o2 + 0p_t_2_o3 - pi_2_1 >= 0
500p_t_3_o1 + 200p_t_3_o2 + 0p_t_3_o3 - pi_3_1 >= 0
0p_t_1_o1 + 0p_t_1_o2 + 0p_t_1_o3 - pi_1_1 - 0p_t_2_o1 - 0p_t_2_o2 -
0p_t_2_o3 +
pi_2_1 >= 0
0p_t_1_o1 + 0p_t_1_o2 + 0p_t_1_o3 - pi_1_1 - 0p_t_3_o1 - 0p_t_3_o2 -
0p_t_3_o3 +
pi_3_1 >= 0
200p_t_2_o1 + 100p_t_2_o2 + 0p_t_2_o3 - pi_2_1 - 200p_t_1_o1 -
100p_t_1_o2 - 0p_
t_1_o3 + pi_1_1 >= 0
200p_t_2_o1 + 100p_t_2_o2 + 0p_t_2_o3 - pi_2_1 - 200p_t_3_o1 -
100p_t_3_o2 - 0p_
t_3_o3 + pi_3_1 >= 0
500p_t_3_o1 + 200p_t_3_o2 + 0p_t_3_o3 - pi_3_1 - 500p_t_1_o1 -
200p_t_1_o2 - 0p_
t_1_o3 + pi_1_1 >= 0
500p_t_3_o1 + 200p_t_3_o2 + 0p_t_3_o3 - pi_3_1 - 500p_t_2_o1 -
200p_t_2_o2 - 0p_
t_2_o3 + pi_2_1 >= 0
bounds
p_t_1_o1 >= 0
p_t_1_o2 >= 0
p_t_1_o3 >= 0
-inf <= pi_1_1 <= +inf
p_t_2_o1 >= 0
p_t_2_o2 >= 0
p_t_2_o3 >= 0
-inf <= pi_2_1 <= +inf
p_t_3_o1 >= 0
p_t_3_o2 >= 0
p_t_3_o3 >= 0
-inf <= pi_3_1 <= +inf
end
```

Fan pays 100   Yinzer pays 400

```
CPLEX> dis sol var -
Variable Name              Solution Value
pi_2_1                        100.000000
pi_3_1                        400.000000
p_t_1_o3                        1.000000
p_t_2_o2                        1.000000
p_t_3_o1                        1.000000
All other variables in the range 1-12 are 0.
```
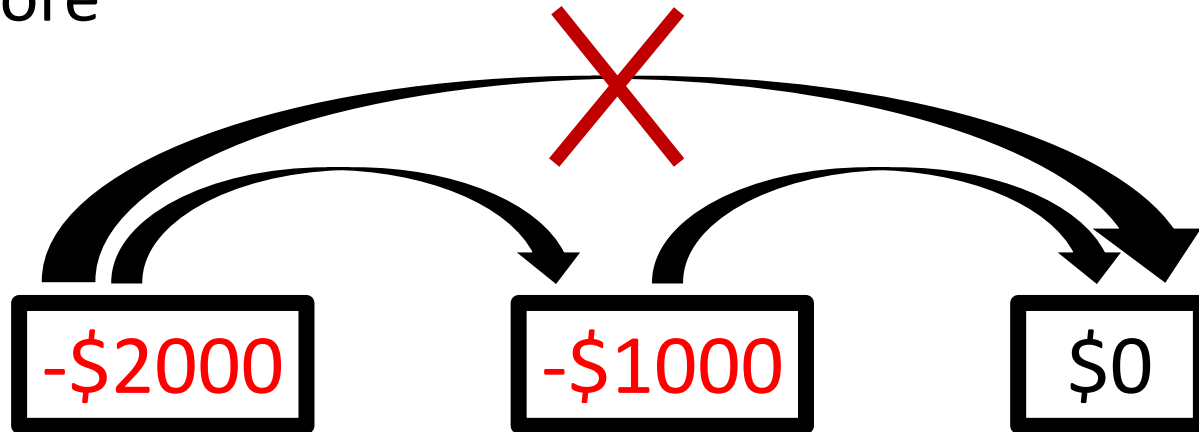
Yinzer gets   Fan gets   Uninterested
Great seat   Decent seat   gets No seat

Failure of the revelation principle with partial verification

- Suppose anyone can secretly borrow another $1000 temporarily, but no more

- Goal: accept people who are (truly) at most $1000 in debt
- Is it possible? Truthfully?

Automated mechanism design – results *when you know the choice function*


with Andrew Kephart
(AAMAS 2015)

		Transfers (T)		No Transfers (NT)	
		Two Outcomes (TO)	*Injective SCF (FI)*	*Two Outcomes (TO)*	*Injective SCF (FI)*
Free Utilities (FU)	*Unrestricted Costs (U)*	NP-c	**NP-c**	NP-c	**NP-c**
	$\{0, \infty\}$ Costs (ZI)	NP-c	**NP-c**	NP-c	P
Targeted Utilities (TU)	*Unrestricted Costs (U)*	**NP-c**	P	NP-c	P
	$\{0, \infty\}$ Costs (ZI)	**NP-c**	P	NP-c	P

Non-bolded results are from:
Auletta, Penna, Persiano, Ventre. Alternatives to truthfulness are hard to recognize. AAMAS 2011

Revelation principle holds with transitivity

- Suppose you can only *overreport* your debt
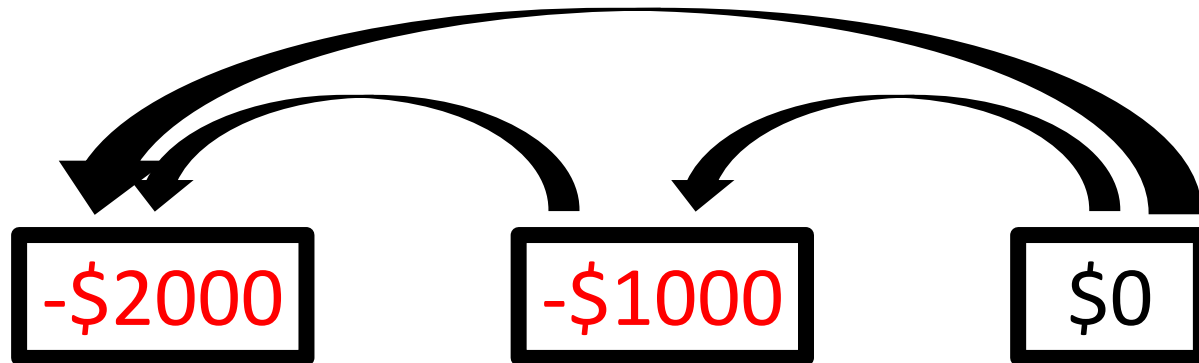
- Goal: accept people who are (truly) at most $1000 in debt
- Is it possible? Truthfully?
- How about: goal: accept people who are (truly) at *least* $1000 in debt
- General conditions under which revelation principle still holds: in Green & Laffont RES '86 and Yu AAMAS '11 (partial verification), and Kephart & C. EC'16 / ACM TEAC'21 (costly signaling)
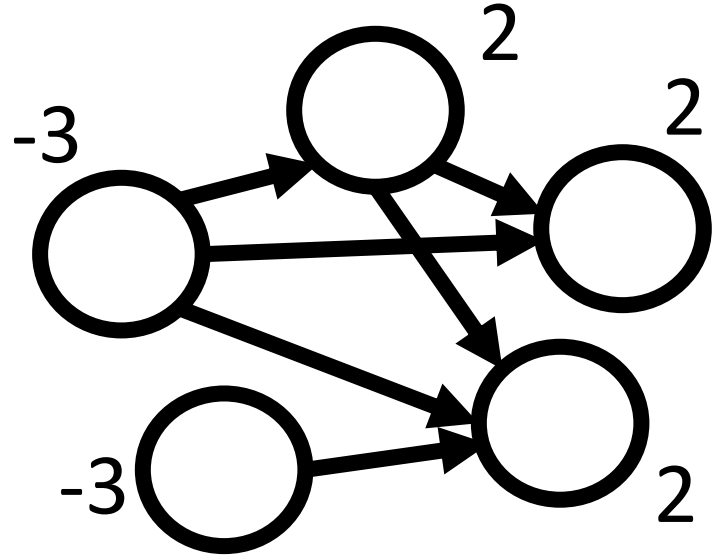
Andrew Kephart

Optimization: reduction to min cut

(when revelation principle holds)

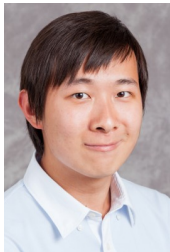types are vertices; edges imply ability
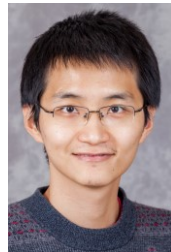to (cost-effectively) misreport

edges between types have capacity ∞

Values are P(type)*value(type)
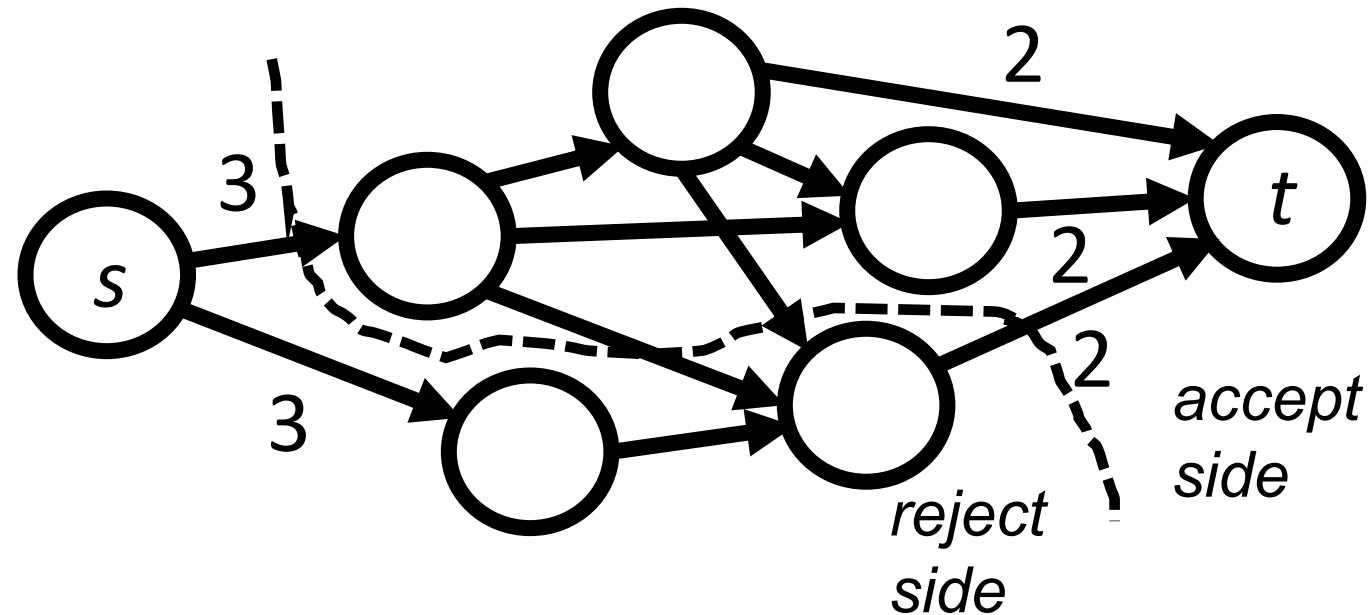
AAAI-21[a], with:

Hanrui Zhang   Yu Cheng

Can be generalized to more outcomes than
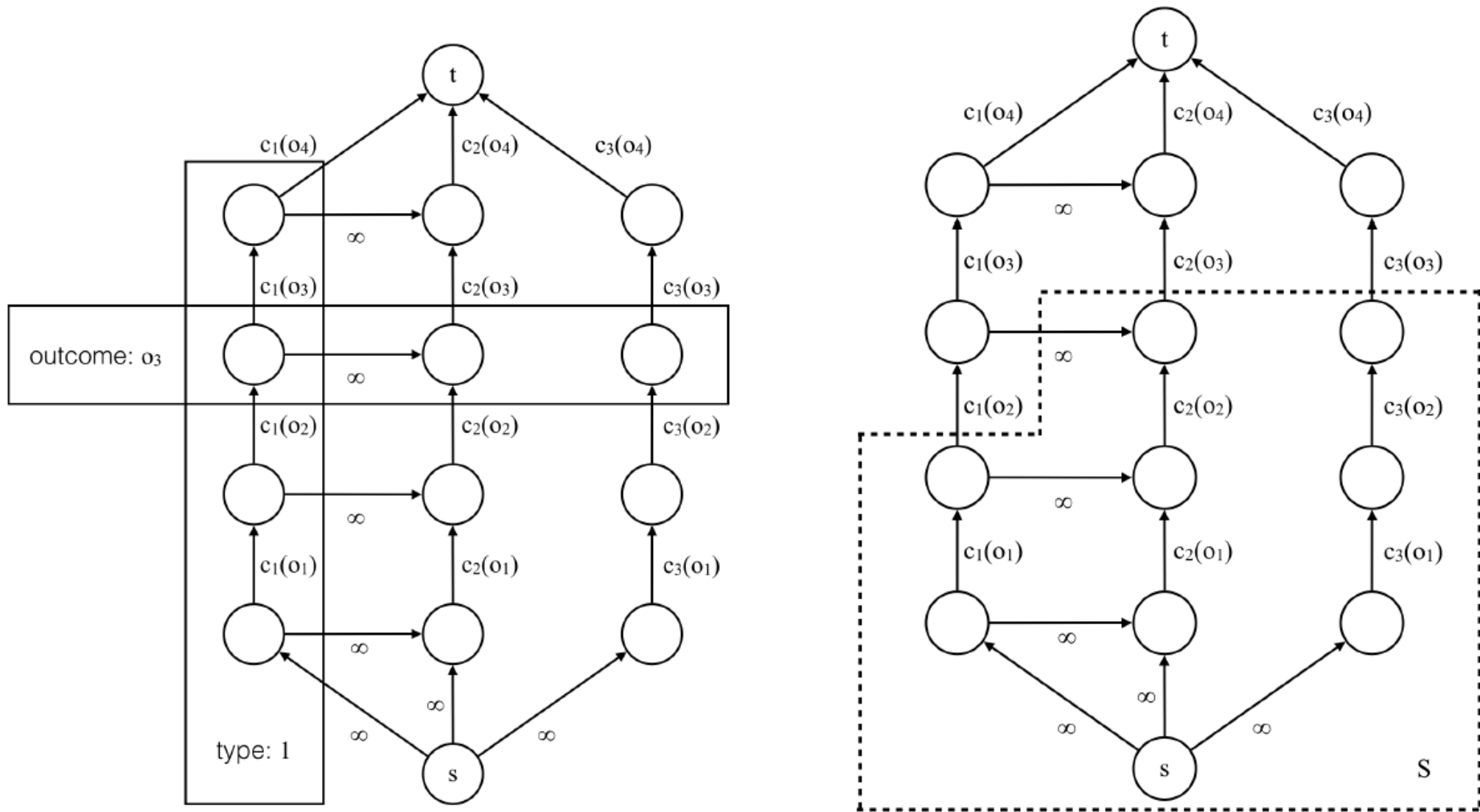accept/reject, **if** types have the same utility
over them.

Figure 1: An example of the graph constructed in Algorithm 1. As highlighted in the left graph, each row corresponds to an outcome and each column corresponds to a type. The horizontal edges with infinite capacity correspond to the fact that type 2 can misreport as type 1. The right graph gives a possible s-t min-cut, which corresponds to a mechanism where $M(1) = o_2$, $M(2) = (o_3)$, and $M(3) = o_3$. The horizontal edges make sure that type 1 never gets a more desirable outcome than type 2, so type 2 never misreports. The cost of the mechanism M is equal to the value of the min-cut, which is $c_1(o_2) + c_2(o_3) + c_3(o_3)$.

Generalization

AAAI-21[b], with:
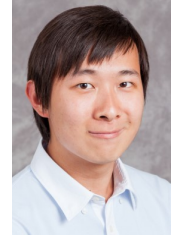
Hanrui Zhang

- considering IC classifiers <u>imposes regularization</u>

- whp for <u>all IC classifiers</u> f in 2^X simultaneously,

$$\hat{\ell_D}(f) = \ell_D(f) \leq \ell_S(f) + O\left(\sqrt{\frac{\text{VC}(X,\rightarrow)}{m}}\right)$$

- $\text{VC}(X, \rightarrow)$: intrinsic dimension of feature space & reporting structure

Intrinsic dimension

- $\text{VC}(X, \rightarrow)$: intrinsic dimension of feature space & reporting structure

  - for any $x, x' \in X$, x <u>can reach</u> x' if there exists a sequence $x = x_1, \ldots, x_k = x'$ such that for all $1 \leq i < k, x_i \rightarrow x_{i+1}$

  - $\text{VC}(X, \rightarrow)$ is the cardinality of the largest $A \subseteq X$, such that for any $x_1, x_2 \in A$ where $x_1 \neq x_2$, x_1 cannot reach x_2

  - in other words, $\text{VC}(X, \rightarrow)$ is the <u>width</u> of the transitive closure of $\rightarrow$

Incentive-compatible classifiers
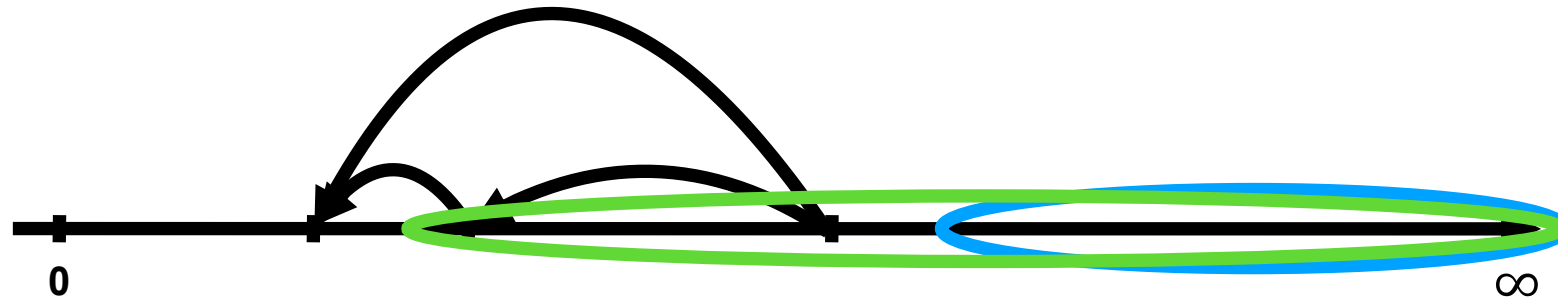
- $X = \mathbb{R}_+, \rightarrow \, = \, \geq, \mathrm{VC}(X, \rightarrow) = 1$

- IC classifiers (e.g., blue and green) = thresholds

    <u>all IC classifiers generalize well</u>

- ERM using efficient algorithm for Bayesian setting discussed earlier
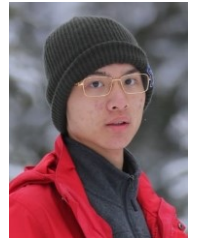
Dropping feature values
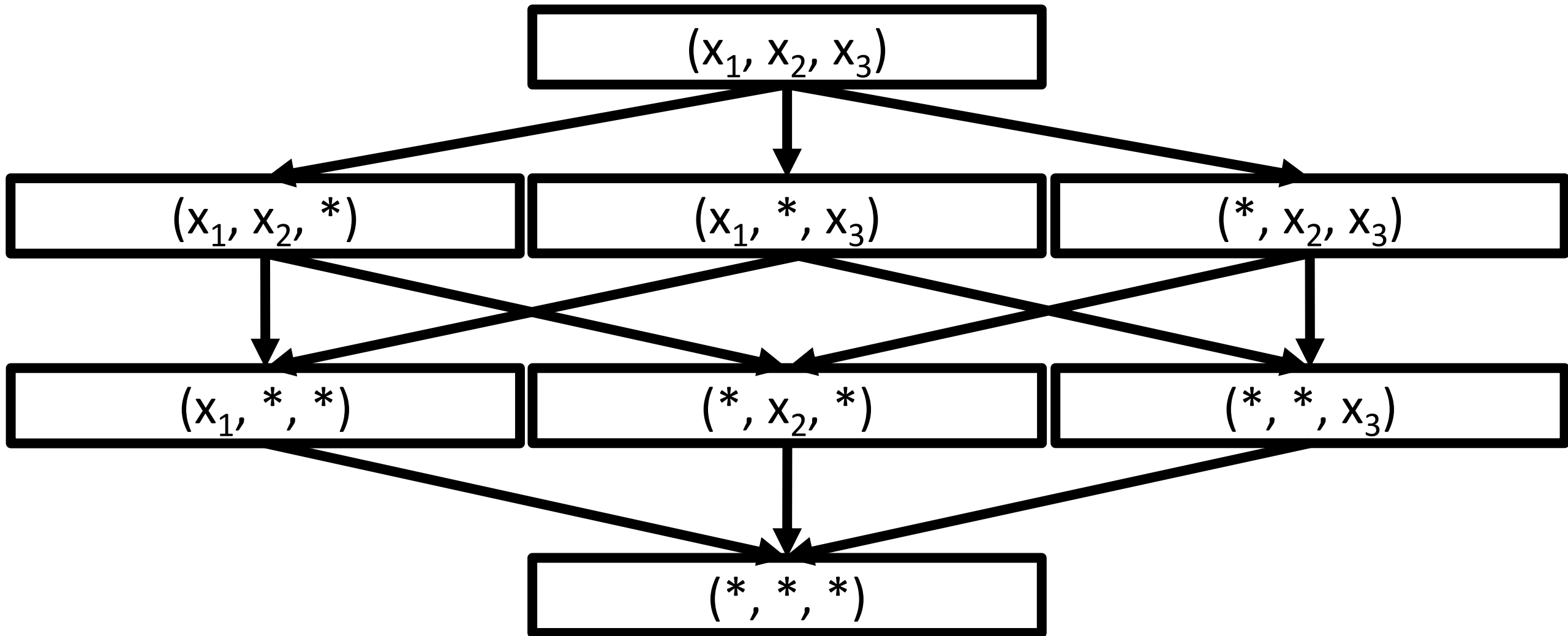
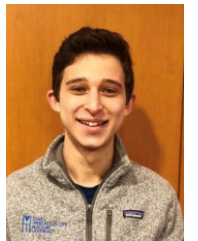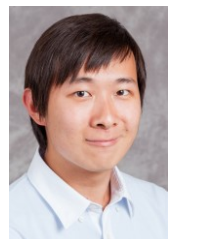# Experimental results: dropping feature values

AAAI-21[c], with:

Anilesh K. Krishnaswamy

Haoming Li

David Rein

Hanrui Zhang

Table 5: Our methods vs. the rest: mean classifier accuracy for $\epsilon = 0.2$, balanced datasets, all features

Classifier	Australia		Germany		Poland		Taiwan	
	Tru.	Str.	Tru.	Str.	Tru.	Str.	Tru.	Str.
HCFS(LR)	.795	.795	.625	.625	.678	.678	.648	.648
HCAPP(LR)	.777	.777	.617	.617	.658	.658	.638	.638
MINCUT	.496	.496	.499	.499	.499	.499	.499	.499
IC-LR	.798	.798	.654	**.654**	.607	.607	.588	.588
HCFS(LR) w/ disc.	.794	.794	.632	.632	.694	.694	.649	.649
HCAPP(LR) w/ disc.	.782	.782	.620	.620	.724	.724	.644	.644
MINCUT w/ disc.	.534	.534	.503	.503	.499	.499	.550	.550
IC-LR w/ disc.	.805	**.805**	.653	.653	.773	**.773**	.667	**.667**
IMP(LR)	.802	.701	**.663**	.523	.729	.507	.657	.501
IMP(LR) w/ disc.	**.809**	.723	.659	.554	**.783**	.503	**.697**	.501

Experimental results: dropping feature values (fewer features)

AAAI-21[c], with:

Anilesh K. Krishnaswamy

Haoming Li

David Rein

Hanrui Zhang

Table 3: Our methods vs. the rest: mean classifier accuracy for $\epsilon = 0.2$, balanced datasets, 4 features

Classifier	Australia		Germany		Poland		Taiwan	
	Tru.	Str.	Tru.	Str.	Tru.	Str.	Tru.	Str.
HC(LR)	.792	**.792**	.639	.639	.659	.659	.648	.648
MINCUT	.770	.770	.580	.580	.501	.501	.652	**.652**
IC-LR	.788	.788	.654	**.654**	.639	.639	.499	.499
IMP(LR)	.796	.791	**.663**	.580	**.714**	**.660**	**.670**	.618
R-F(LR)	**.808**	.545	.631	.508	.670	.511	.665	.590

Table 4: Our methods vs. the rest: mean classifier accuracy for $\epsilon = 0.2$, balanced datasets, 4 features ("w/ disc." stands for "with discretization of features")

| Classifier	Australia		Germany		Poland		Taiwan	
	Tru.	Str.	Tru.	Str.	Tru.	Str.	Tru.	Str.
HC(LR) w/ disc.	.794	.794	.641	.641	.692	.692	.650	**.650**
MINCUT w/ disc.	.789	.789	.629	.629	.692	.692	.649	.649
IC-LR w/ disc.	**.800**	**.800**	.651	**.651**	.698	**.698**	.646	.646
IMP(LR) w/ disc.	.799	.762	**.652**	.577	**.719**	.631	**.686**	.541
R-F(LR) w/ disc.	.796	.542	.633	.516	.708	.522	.684	.587
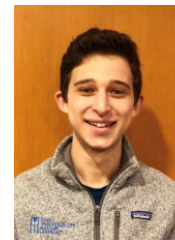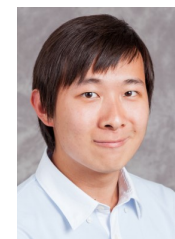
Hillclimbing and the hierarchy

associate classifier with each node in the hierarchy
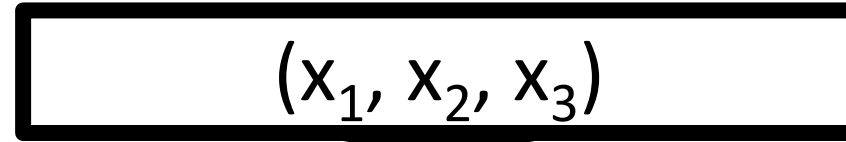
agent is accepted if it is accepted by any one of the classifiers it can access

f_{123} (x_1, x_2, x_3)

f_{12*} $(x_1, x_2, *)$ $(x_1, *, x_3)$ $(*, x_2, x_3)$

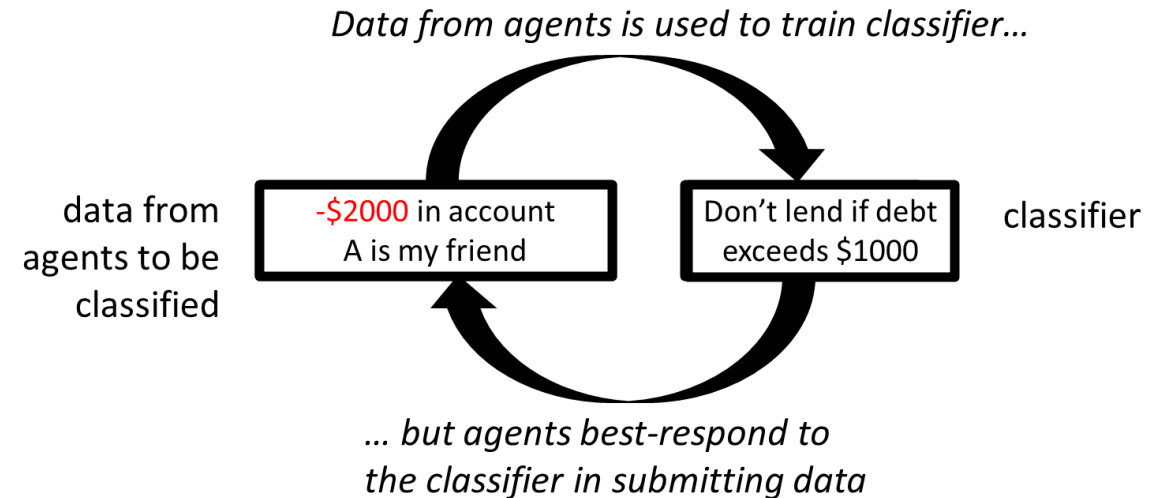$(x_1, *, *)$ $(*, x_2, *)$ $(*, *, x_3)$

$(*, *, *)$
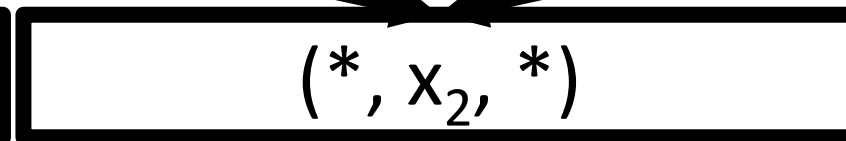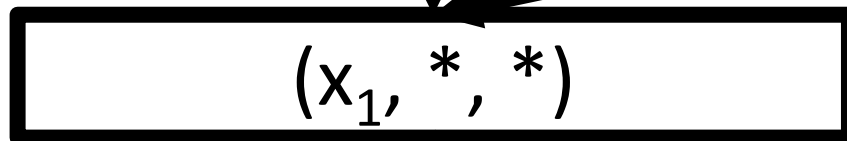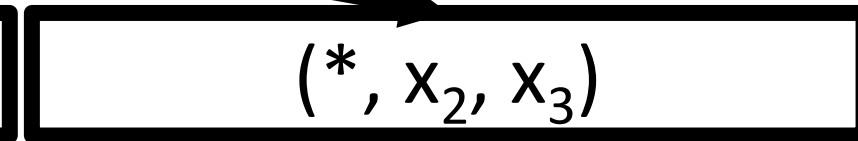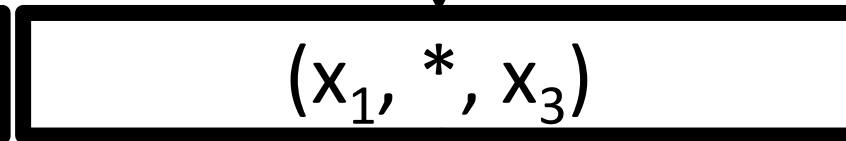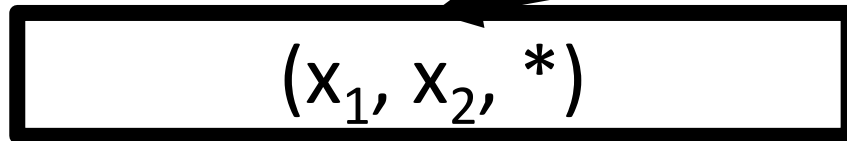
this is without loss of generality

HillClimbing: repeatedly retrain some node's classifier taking into account all examples that can access it and are rejected elsewhere

Future research

- What if agents' effort can change their type? [see also Kleinberg and Raghavan 2019]

- Can we use standard ML methods in a black-box way?

- Truly online models without separate training stage on trusted data

Data from agents is used to train classifier…

data from agents to be classified

| -$2000 in account A is my friend | Don't lend if debt exceeds $1000 |

classifier

… but agents best-respond to the classifier in submitting data

THANK YOU FOR YOUR ATTENTION!