Automated Mechanism Design for Strategic Classification

Vincent Conitzer, joint work with:

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?

In depth: Want a loan? China's tech giants are at your service

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

AI in Dating Apps: The Changing Face of Online Dating Industry
“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]
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. [...]
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.
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.

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."
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

Data from agents is used to train 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://giphy.com/gifs/heute-show-yes-celebration-3o75279fBGgbO848ZW
Models of (mis)reporting: direct revelation
Agent’s type = feature values

- $2000 in account
  A and B are my friends

- $2000 in account
  A, B, and C are my friends

$0 in account
A and B are my friends

$0 in account
A, B, and C are my friends
Interlude: Mechanism design for traditional applications

Selling tickets to a Steelers game

- **Uninterested**
- **Fan**
- **true Yinzer**

**Incentive compatible:** No type benefits from misreporting

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

- **Decent**
  - \( v_U(G) = v_U(D) = v_U(N) = 0 \)
  - \( v_F(G) = 200, \quad v_F(D) = 100, \quad v_F(N) = 0 \)
  - \( v_Y(G) = 500, \quad v_Y(D) = 200, \quad v_Y(N) = 0 \)

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

<table>
<thead>
<tr>
<th></th>
<th>unlimited misreporting</th>
<th>partial verification / costly misreporting</th>
</tr>
</thead>
<tbody>
<tr>
<td>identical preferences</td>
<td>trivial / can’t do much</td>
<td>some classification settings</td>
</tr>
<tr>
<td>distinct preferences</td>
<td>traditional applications</td>
<td>other classification settings</td>
</tr>
</tbody>
</table>

Nobel Prizes in Economics:
- 2007 (mechanism design): Hurwicz, Maskin, Myerson
- 2012 (matching mechanisms): Roth, Shapley
- 2020 (auction mechanisms): 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 \[\text{[C. \\& Sandholm UAI 2002 and subsequent work]}\]

-- example

**INPUT**

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

**OUTPUT**

• Probability distribution: .3U, .4F, .3Y
• Other details: objective (revenue), randomization allowed (yes), ...

• 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
0.3\pi_1_1 + 0.4\pi_2_1 + 0.3\pi_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 + p_t_2_o1 = 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 + p_t_3_o1 = 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

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.

Fan pays 100
Yinzer pays 400

Yinzer gets Great seat
Fan gets Decent seat
Uninterested 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*

<table>
<thead>
<tr>
<th>Free Utilities (FU)</th>
<th>Unrestricted Costs (U)</th>
<th>NP-c</th>
<th>NP-c</th>
<th>NP-c</th>
<th>NP-c</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>{0, ∞} Costs (ZI)</td>
<td>NP-c</td>
<td>NP-c</td>
<td>NP-c</td>
<td>P</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Targeted Utilities (TU)</th>
<th>Unrestricted Costs (U)</th>
<th>NP-c</th>
<th>P</th>
<th>NP-c</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>{0, ∞} Costs (ZI)</td>
<td>P</td>
<td>P</td>
<td>NP-c</td>
<td>P</td>
</tr>
</tbody>
</table>

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

with Andrew Kephart (AAMAS 2015)
Revelation principle holds with transitivity

• Suppose you can only *overreport* your debt

- $2000  
- $1000  
  $0

• 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)
Optimization: reduction to min cut

(when revelation principle holds)

Values are $P(\text{type}) \times \text{value(\text{type})}$

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

• considering IC classifiers imposes regularization

• whp for all IC classifiers $f$ in $2^X$ simultaneously,

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

• $\text{VC}(X,\to)$: intrinsic dimension of feature space & reporting structure
Intrinsic dimension

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

- for any $x, x' \in X$, $x$ can reach $x'$ if there exists a sequence $x = x_1, ..., x_k = x'$ 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 width of the transitive closure of $\rightarrow$
Incentive-compatible classifiers

- \( X = \mathbb{R}_+, \rightarrow \geq, VC(X, \rightarrow) = 1 \)

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

  all IC classifiers generalize well

- ERM using efficient algorithm for Bayesian setting discussed earlier
Dropping feature values

\[(x_1, x_2, x_3)\]

\[(x_1, x_2, *)\]  \[(x_1, *, x_3)\]  \[(*, x_2, x_3)\]

\[(x_1, *, *)\]  \[(*, x_2, *)\]  \[(*, *, x_3)\]

\[(*, *, *)\]
Experimental results: dropping feature values

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

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Australia</th>
<th></th>
<th>Germany</th>
<th></th>
<th>Poland</th>
<th></th>
<th>Taiwan</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>True</td>
<td>Str.</td>
<td>True</td>
<td>Str.</td>
<td>True</td>
<td>Str.</td>
<td>True</td>
</tr>
<tr>
<td>HCFS(LR)</td>
<td>.795</td>
<td>.795</td>
<td>.625</td>
<td>.625</td>
<td>.678</td>
<td>.678</td>
<td>.648</td>
<td>.648</td>
</tr>
<tr>
<td>HCAPP(LR)</td>
<td>.777</td>
<td>.777</td>
<td>.617</td>
<td>.617</td>
<td>.658</td>
<td>.658</td>
<td>.638</td>
<td>.638</td>
</tr>
<tr>
<td>MINCUT</td>
<td>.496</td>
<td>.496</td>
<td>.499</td>
<td>.499</td>
<td>.499</td>
<td>.499</td>
<td>.499</td>
<td>.499</td>
</tr>
<tr>
<td>IC-LR</td>
<td>.798</td>
<td>.798</td>
<td>.654</td>
<td></td>
<td>.607</td>
<td>.607</td>
<td>.588</td>
<td>.588</td>
</tr>
<tr>
<td>HCFS(LR) w/ disc.</td>
<td>.794</td>
<td>.794</td>
<td>.632</td>
<td>.632</td>
<td>.694</td>
<td>.694</td>
<td>.649</td>
<td>.649</td>
</tr>
<tr>
<td>HCAPP(LR) w/ disc.</td>
<td>.782</td>
<td>.782</td>
<td>.620</td>
<td>.620</td>
<td>.724</td>
<td>.724</td>
<td>.644</td>
<td>.644</td>
</tr>
<tr>
<td>MINCUT w/ disc.</td>
<td>.534</td>
<td>.534</td>
<td>.503</td>
<td>.503</td>
<td>.499</td>
<td>.499</td>
<td>.550</td>
<td>.550</td>
</tr>
<tr>
<td>IC-LR w/ disc.</td>
<td>.805</td>
<td></td>
<td>.653</td>
<td>.653</td>
<td>.773</td>
<td>.773</td>
<td>.667</td>
<td>.667</td>
</tr>
<tr>
<td>IMP(LR)</td>
<td>.802</td>
<td>.701</td>
<td>.663</td>
<td>.523</td>
<td>.729</td>
<td>.507</td>
<td>.657</td>
<td>.501</td>
</tr>
<tr>
<td>IMP(LR) w/ disc.</td>
<td>.809</td>
<td>.723</td>
<td>.659</td>
<td>.554</td>
<td>.783</td>
<td>.503</td>
<td>.697</td>
<td>.501</td>
</tr>
</tbody>
</table>
Experimental results: dropping feature values (fewer features)

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

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Australia</th>
<th>Germany</th>
<th>Poland</th>
<th>Taiwan</th>
</tr>
</thead>
<tbody>
<tr>
<td>HC(LR)</td>
<td>.792</td>
<td><strong>.792</strong></td>
<td>.639</td>
<td>.639</td>
</tr>
<tr>
<td>IC-LR</td>
<td>.788</td>
<td>.788</td>
<td>.654</td>
<td><strong>.654</strong></td>
</tr>
<tr>
<td>Imp(LR)</td>
<td>.796</td>
<td>.791</td>
<td><strong>.663</strong></td>
<td>.580</td>
</tr>
<tr>
<td>R-F(LR)</td>
<td><strong>.808</strong></td>
<td>.545</td>
<td>.631</td>
<td>.508</td>
</tr>
</tbody>
</table>

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")

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Australia</th>
<th>Germany</th>
<th>Poland</th>
<th>Taiwan</th>
</tr>
</thead>
<tbody>
<tr>
<td>HC(LR) w/ disc.</td>
<td>.794</td>
<td>.794</td>
<td>.641</td>
<td>.641</td>
</tr>
<tr>
<td>Mincut w/ disc.</td>
<td>.789</td>
<td>.789</td>
<td>.629</td>
<td>.629</td>
</tr>
<tr>
<td>IC-LR w/ disc.</td>
<td><strong>.800</strong></td>
<td><strong>.800</strong></td>
<td>.651</td>
<td><strong>.651</strong></td>
</tr>
<tr>
<td>Imp(LR) w/ disc.</td>
<td>.799</td>
<td>.762</td>
<td><strong>.652</strong></td>
<td>.577</td>
</tr>
<tr>
<td>R-F(LR) w/ disc.</td>
<td>.796</td>
<td>.542</td>
<td>.633</td>
<td>.516</td>
</tr>
</tbody>
</table>
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

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

(this is without loss of generality)
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

THANK YOU FOR YOUR ATTENTION!