An Alternative View on AI: Collaborative Learning, Incentives, and Social Welfare

Michael I. Jordan

University of California, Berkeley
Perspectives on AI

• The classical “human-imitative” perspective
  - cf. AI in the movies

• The “intelligence augmentation” (IA) perspective
  - cf. search engines, recommendation systems, natural language translation
  - the system need not be intelligent itself, but it reveals patterns that humans can make use of

• The “intelligent infrastructure” (II) perspective
  - cf. transportation, intelligent dwellings, urban planning
  - large-scale, distributed collections of data flows and loosely-coupled decisions
  - novel market mechanisms and novel deliberative mechanisms based on data flows
The 1950s AI Perspective

• A goal of understanding the intelligence of an individual human and building computers that mimic such intelligence
  • and possibly improve on it
• Not very clear what the overall engineering goal is
  • what kind of systems will such intelligences be embedded in
  • what kind of problems will such systems solve?
  • seems naïve to expect to solve real-world problems---in domains such as health care, climate change, commerce, etc---with such a vague premise
A Counterpoint

• Intelligence is as much about the collective as it is about the individual

• In terms of establishing goals for an emerging engineering field, thinking in terms of collectives seems at least as urgent and promising as thinking in terms of individual intelligence

• There may be new forms of collectives that can emerge if we put our minds to it
Federated Learning Paradigm

Purported to aim at collective mechanisms, but does it?
Data, Creators, Values, and Collaborations

• In real life, the “nodes” are often people, and their data is not something to simply be streamed and aggregated

• People often value their data

• They may wish to reveal aspects of their data if (and only if) they obtain commensurate benefits

• One way to start to understand this is to develop blends of microeconomics and machine learning

• Learning-aware mechanisms and mechanism-aware learning
Music in the Data Age

• Use data to structure a two-sided market; e.g., by providing a **dashboard** to musicians, letting them learn where their audience is
  - the musician can give shows where they have an audience

• I.e., consumers and producers become linked, and value flows: a market is created
  - the company that creates this market profits simply by taking a cut from the transactions

• Bring in brands and create a three-way market
  - the brands can partner with specific musicians based on affinities

• The company **United Masters** is doing precisely this; [www.unitedmasters.com](http://www.unitedmasters.com)
Some Problems at the Interface of ML and Econ

• Relationships among optima, equilibria, and dynamics
• Exploration, exploitation, and incentives in multi-way markets
• Information asymmetries, contracts and statistical inference
• Strategic classification
• Uncertainty quantification for black box and adversarial settings
• Calibrating predictions for inference and decision-making
• Mechanism design with learned preferences
Statistical Contract Theory

Stephen Bates  Michael Sklar  Jake Soloff
The Theory of Incentives

- **Contract theory** is one branch of the theory of incentives (auction theory is another branch)

- In contract theory, agents possess private information and a principal wishes to incentivize them to take actions that depend on that private information
  - the goal is overall social welfare, or revenue

- For example, services such as airlines have “business fares” and “economy fares”
  - this allows them to offer different prices to agents who have different willingness to pay, without requiring agents to reveal their private values

- The design problem is to determine a menu of options, of the form (service, price), from which agents select
Clinical Trials

Average Cost of Clinical Trial

Department of Health and Human Services, 2014

<table>
<thead>
<tr>
<th>Therapeutic Area</th>
<th>Phase 1</th>
<th>Phase 2</th>
<th>Phase 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anti-Infective</td>
<td>$4.2 (5)</td>
<td>$14.2 (6)</td>
<td>$22.8 (5)</td>
</tr>
<tr>
<td>Cardiovascular</td>
<td>$2.2 (9)</td>
<td>$7.0 (13)</td>
<td>$25.2 (3)</td>
</tr>
<tr>
<td>Central Nervous System</td>
<td>$3.9 (6)</td>
<td>$13.9 (7)</td>
<td>$19.2 (7)</td>
</tr>
<tr>
<td>Dermatology</td>
<td>$1.8 (10)</td>
<td>$8.9 (12)</td>
<td>$11.5 (13)</td>
</tr>
<tr>
<td>Endocrine</td>
<td>$1.4 (12)</td>
<td>$12.1 (10)</td>
<td>$17.0 (9)</td>
</tr>
<tr>
<td>Gastrointestinal</td>
<td>$2.4 (8)</td>
<td>$15.8 (4)</td>
<td>$14.5 (11)</td>
</tr>
<tr>
<td>Genitourinary System</td>
<td>$3.1 (7)</td>
<td>$14.6 (5)</td>
<td>$17.5 (8)</td>
</tr>
<tr>
<td>Hematology</td>
<td>$1.7 (11)</td>
<td>$19.6 (1)</td>
<td>$15.0 (10)</td>
</tr>
<tr>
<td>Immunomodulation</td>
<td>$6.6 (1)</td>
<td>$16.0 (3)</td>
<td>$11.9 (12)</td>
</tr>
<tr>
<td>Oncology</td>
<td>$4.5 (4)</td>
<td>$11.2 (11)</td>
<td>$22.1 (6)</td>
</tr>
<tr>
<td>Ophthalmology</td>
<td>$5.3 (2)</td>
<td>$13.8 (8)</td>
<td>$30.7 (2)</td>
</tr>
<tr>
<td>Pain and Anesthesia</td>
<td>$1.4 (13)</td>
<td>$17.0 (2)</td>
<td>$52.9 (1)</td>
</tr>
<tr>
<td>Respiratory System</td>
<td>$5.2 (3)</td>
<td>$12.2 (9)</td>
<td>$23.1 (4)</td>
</tr>
</tbody>
</table>

(in millions of dollars)

Immense social investment in clinical trials
Contract Theory

principal

• Has only partial knowledge
• Must incentivize the agents

agent

• Has private information
• Strategic and self-interested
How Should the FDA Test?

<table>
<thead>
<tr>
<th>type</th>
<th>P(approve)</th>
<th>P(non-approve)</th>
</tr>
</thead>
<tbody>
<tr>
<td>bad drugs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\theta = 0$</td>
<td>0.05</td>
<td>0.95</td>
</tr>
<tr>
<td>good drugs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\theta = 1$</td>
<td>0.80</td>
<td>0.20</td>
</tr>
</tbody>
</table>

(5% type-1 error)
(80% power)

Is this a good protocol?

**Case 1: small profit.** $20$ million cost to run trial. $200$ million if approved.

\[
\mathbb{E}[\text{profit}|\theta = 0] = -10 \text{ million}
\]

All approvals are good drugs!

**Case 2: large profit.** $20$ million cost to run trial. $2$ billion if approved.

\[
\mathbb{E}[\text{profit}|\theta = 0] = 80 \text{ million}
\]

Many bad drugs are approved!
Statistical Contracts

Denote the agent’s private information as $\theta \in \Theta$

Present the agent with the following opt-in protocol:

1. Agent pays $R$

2. Agent chooses payout function $f$ from menu $\mathcal{F}$

3. Statistical trial yields random variable $Z \sim P_\theta$

4. Agent receives payoff $f(Z)$
   Principal receives utility $u(\theta, f(Z))$

Agent acts to maximize their payoff: $f^{br} = \arg\max_{f \in \mathcal{F}} \mathbb{E}_{Z \sim P_\theta}[f(Z)]$

our task: design this menu
Incentive Alignment

null agents: $\Theta_0 \subset \Theta$ \quad $u(\theta_0, f(Z)) \leq 0$, decreasing in $f(Z)$ for $\theta_0 \in \Theta_0$

nonnull agents: $\Theta \setminus \Theta_0$ \quad $u(\theta_1, f(Z)) \geq 0$, increasing in $f(Z)$ for $\theta_1 \notin \Theta_0$

The principal wants to transact as much as possible with good agents

**Definition** (Incentive-aligned contract)

A menu $\mathcal{F}$ is *incentive-aligned* if for all $f \in \mathcal{F}$ and $\theta_0 \in \Theta_0$

$$\mathbb{E}_{Z \sim P_{\theta_0}} [f(Z) - R] \leq 0$$

agent’s expected profit

\textbf{note: $p \leq .05$ protocol not incentive aligned}

On average, null drugs are not profitable, so null agents are incentivized to drop out
E-values: Statistical Evidence on the Right Scale

Definition

A random variable \( X \geq 0 \) is an E-value for null hypothesis \( \Theta_0 \) if for all \( \theta_0 \in \Theta_0 \)

\[
E_{Z \sim P_{\theta_0}}[X] \leq 1
\]

Theorem

A contract is incentive-aligned if and only if all payoff functions are E-values.
Incentivizing Data Sharing in Federated Learning

• Multiple agents cooperate with each other and with a principal to build a better statistical model than anyone could do unilaterally
  • mostly this literature has developed without considering incentives
  • free riding is a practical concern

• We adapt our statistical contract theory perspective to the problem
  • we design an incentive-compatible mechanism that incentivizes agents to contribute a maximum amount of data (rather than eliciting private types)
  • a key tool is statistical accuracy shaping

Prediction-Powered Inference

Anastasios Angelopoulos
Stephen Bates
Clara Fannjiang
Tijana Zrnic
PROTEIN POWER

AI network predicts highly accurate 3D structures for the human proteome

Troubled waters: The race to save the Great Barrier Reef from climate change
Coronavirus: Tracing back to find the origins of SARS-CoV-2
Storage hunting: Quantifying carbon held in Africa's montane forests

Method of the Year 2021: Protein structure prediction
Protein structure studies

Hundreds of millions of amino acid sequences with protein structures predicted by AlphaFold

Hundreds of thousands of amino acid sequences with protein structures from X-ray crystallography

Goal: correlate sequence information with structural information
The importance of intrinsic disorder for protein phosphorylation

Lilia M. Iakoucheva, Predrag Radivojac, Celeste J. Brown, Timothy R. O'Connor, Jason G. Sikes, Zoran Obradovic and A. Keith Dunker
The importance of intrinsic disorder for protein phosphorylation

Lilia M. Iakoucheva, Predrag Radivojac¹, Celeste J. Brown, Timothy R. O’Connor, Jason G. Sikes, Zoran Obradovic¹ and A. Keith Dunker*
The importance of intrinsic disorder for protein phosphorylation
Lilia M. Iakoucheva, Predrag Radivojac¹, Celeste J. Brown, Timothy R. O'Connor, Jason G. Sikes, Zoran Obradovic¹ and A. Keith Dunker* 

2004
Not enough structures overlapping with post-translational modification (PTM) data.

The structural context of posttranslational modifications at a proteome-wide scale
Isabell Bludau¹, Sander Willems¹, Wen-Feng Zeng¹, Maximilian T. Strauss², Fynn M. Hansen¹, Maria C. Tanzer¹, Ozge Karayel¹, Brenda A. Schulman³, Matthias Mann¹,²* 

2022
Quantify association between PTMs and IDR by computing:

\[
\text{oDDS ratio} = \frac{\mathbb{P}(\text{IDR} \mid \text{PTM})}{\mathbb{P}(\text{IDR} \mid \text{no PTM})}
\]

2022
200+ million AlphaFold-predicted structures

2004
~10k experimental structures in PDB
The importance of intrinsic disorder for protein phosphorylation
Lilia M. Iakoucheva, Predrag Radivojac¹, Celeste J. Brown, Timothy R. O’Connor,
Jason G. Sikes, Zoran Obadovíc and A. Keith Dunker*  

2002
Not enough structures overlapping with post-translational modification (PTM) data.

METHODS AND RESOURCES
PLOS BIOLOGY
Published: May 16, 2022

The structural context of posttranslational modifications at a proteome-wide scale
Isabell Bludau¹, Sander Willems¹, Wen-Feng Zeng¹, Maximilian T. Strauss², Fynn M. Hansen¹, Maria C. Tanzer¹, Özge Karayel¹, Brenda A. Schulman³, Matthias Mann¹,²*  

2022
Quantify association between PTMs and IDRs by computing:

\[
\text{odds ratio} = \frac{\mathbb{P}(\text{IDR} \mid \text{PTM})}{\mathbb{P}(\text{IDR} \mid \text{no PTM})}
\]  

2004
200+ million AlphaFold-predicted structures

~10k experimental structures in PDB
Predictions are being used for scientific inquiry.

The importance of intrinsic disorder for protein phosphorylation

Lilia M. Iakoucheva, Predrag Radivojac¹, Celeste J. Brown, Timothy R. O’Connor,
Jason G. Sikes, Zoran Obradovic¹ and A. Keith Dunker*

Not enough structures overlapping with post-translational modification (PTM) data.

The structural context of posttranslational modifications at a proteome-wide scale

Isabell Bludau¹, Sander Willems¹, Wen-Feng Zeng¹, Maximilian T. Strauss², Fynn
M. Hansen¹, Maria C. Tanzer¹, Ozge Karayel¹, Brenda A. Schulman³, Matthias Mann¹.²+*
Predictions are being used for scientific inquiry.

Disease variant prediction with deep generative models of evolutionary data

Quantify association between PTMs and IDRs by computing the odds ratio:

\[
\frac{\mathbb{P}(\text{IDR} \mid \text{PTM})}{\mathbb{P}(\text{IDR} \mid \text{no PTM})}
\]

The structural context of posttranslational modifications at a proteome-wide scale

2022 predicted IDRs

Quantify association between PTMs and IDRs by computing:

predicted IDRs

odds ratio
Prediction-powered inference

\[ \frac{\mathbb{P}(\text{intrinsic disorder} \mid \text{PTM})}{\mathbb{P}(\text{intrinsic disorder} \mid \text{no PTM})} \]

odds ratio
**Prediction-powered inference: problem setting**

<table>
<thead>
<tr>
<th>labeled data</th>
<th>unlabeled data</th>
</tr>
</thead>
<tbody>
<tr>
<td>(X)</td>
<td>(X')</td>
</tr>
<tr>
<td>(Y)</td>
<td>(Y') (unobserved)</td>
</tr>
<tr>
<td>predictions</td>
<td></td>
</tr>
<tr>
<td>(f = f(X))</td>
<td>(f' = f(X'))</td>
</tr>
</tbody>
</table>

Estimand of interest (mean, quantile, regression coefficient, etc.): \(\theta^*\)

Goal: construct confidence set, \(C_{\alpha}^{\text{PP}}\), that are **valid**:

\[
P(\theta^* \in C_{\alpha}^{\text{PP}}) \geq 1 - \alpha
\]

**classical** approach

- use only labeled data
- valid, but lose out on information from abundant predictions

**imputed** approach

- treat predictions as gold-standard labels
- abundant predictions, but **invalid** because predictions can contain systematic errors
Prediction-powered inference: problem setting

\[
\begin{align*}
X & \quad X' \\
Y & \quad Y' \text{(unobserved)} \\
f = f(X) & \quad f' = f(X')
\end{align*}
\]

Estimand of interest (mean, quantile, regression coefficient, etc.): \( \theta^* \)

Goal: construct confidence set, \( C_{\alpha}^{\text{PP}} \), that are valid:
\[
P(\theta^* \in C_{\alpha}^{\text{PP}}) \geq 1 - \alpha
\]

**classical** approach

use only labeled data
valid, but lose out on information from abundant predictions

**imputed** approach

treat predictions as gold-standard labels
abundant predictions, but **invalid** because predictions can contain systematic errors

We want the best of both worlds.
Gene expression

• Want to estimate median gene expression level with differing *promoters* (regulatory DNA)

  (Vaishnav et. al. *Nature* ‘22)

• Predictive model: transformer developed in Vaishnav et. al.
California census

- 2018 CA census data

- Estimand: logistic regression coefficient of income when predicting whether person has private health insurance

- Boosting model based on ten other covariates
1. Identify Rectifier
The rectifier, $\Delta^f$, is a estimand-specific notion of error.
We give a general recipe for identifying the rectifier.

2. Confidence Set on Rectifier
Use the labeled data to construct a confidence set, $R$, for the rectifier.

3. Prediction-Powered Confidence Set
Construct $C_{PP}$ by including all possible rectified values of $\theta^f$.

Principle of prediction-powered inference

For the mean value of $Y$:
rectifier is the bias

$$\Delta^f = \mathbb{E}[f - Y]$$

$$\tilde{\theta}^f = \mathbb{E}[f]$$

$$= \mathbb{E}[Y]$$
Convex Estimation Problems

\[ \theta^* = \arg\min_{\theta} \mathbb{E}[\ell_\theta(X, Y)] \quad \text{e.g. mean, median, quantiles; linear, logistic regression coefficients} \]

gradient of loss \( g_\theta(X, Y) \equiv \frac{\partial}{\partial \theta} \ell_\theta(X, Y) \)

Build confidence set that contains \( \theta^* \): the value of \( \theta \) such that \( \mathbb{E}[g_\theta(X, Y)] = 0 \).

\[ \mathbb{E}[g_\theta(X, f)] - \mathbb{E}[(g_\theta(X, f) - g_\theta(X, Y))] = 0 \]

rectifier \( \Delta^f_\theta \)

estimate using only predictions

\[ \mathbb{E}[g_\theta(X, f)] - \mathbb{E}[(g_\theta(X, f) - g_\theta(X, Y))] = 0 \]

build confidence set \( R_\theta \) for rectifier

using labeled data: \( g_\theta(X_i, f_i) - g_\theta(X_i, Y_i) \)

**Theorem.** Take \( C^{PP} = \{ \theta : 0 \in \mathbb{E}[g_\theta(X, f)] - R_\theta \} \), where for each \( \theta \), the confidence set \( R_\theta \) contains the rectifier \( \Delta^f_\theta \) with probability at least \( 1 - \alpha \). Then, \( C^{PP} \) is valid:

\[ \mathbb{P}(\theta^* \in C^{PP}) \geq 1 - \alpha. \]
A Personal View on “AI”

• It reflects the emergence of a new engineering field, embodied in large-scale systems that link humans in new ways

• Cf. chemical engineering in the 40s and 50s
  - built on chemistry, fluid mechanics, etc
  - driven by the possibility of building chemical factories

• Cf. electrical engineering in the late 19th century
  - built on electromagnetism, optics, etc
  - clear goals in terms of human welfare

• The new field builds on inferential ideas, algorithmic ideas, and economic ideas from the past three centuries

• But its emergence is being warped by being cast in terms of poorly thought-through, naïve, old-style AI aspirations
Three Foundational Disciplines

- Statistics
  - econometrics
  - machine learning
- Economics
  - algorithmic game theory
- Computer Science
Some Further Reading

(see www.cs.berkeley.edu/~jordan/publications.html)

