Machine learning for causality?

@kordinglab
The bitter Lesson (Sutton, 2019)

- Clever human solutions are eventually beat by general purpose machine learning
  - Chess
  - Go
  - Speech recognition
  - Image recognition
  - Natural Language Processing

The Bitter Lesson

Rich Sutton

March 13, 2019

The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin. The ultimate reason for this is Moore's law, or rather its generalization of continued exponentially falling cost per unit of computation. Most AI research has been conducted as if the computation available to the agent were constant (in which case leveraging human knowledge would be one of the only ways to improve performance) but, over a slightly longer time than a typical research project, massively more computation inevitably
How good we are not after ML but after causality

- Medicine is almost always after causality
- And ML merely models correlations
3 Schools

- **RCT or bust**
  - Randomize, avoid threads
  - Very strong strategy for large audience, high value
  - Weak for underserved populations
  - Gold standard (also C. elegans)

- **Observational**
  - Fit complex models, Correct for various variables
  - Lack of causal validity

- **Quasiexperimental**
  - Find a place where the world contains randomization
  - Usually not possible
Intuition: Simulate a trivial causal system

\[ \tilde{x}_{t+1} = \sigma(A\tilde{x}_t + \epsilon_t) \]

- \( \tilde{x}_t \) is an \( n \)-dimensional vector representing our \( n \)-neuron system at timestep \( t \)
- \( \sigma \) is a sigmoid nonlinearity
- \( A \) is our \( n \times n \) causal ground truth connectivity matrix (more on this later)
- \( \epsilon_t \) is random noise: \( \epsilon_t \sim N(0, I_n) \)
- \( \tilde{x}_0 \) is initialized to \( 0 \)

Is correlation \( \sim \) causation?
Great in small system
Not so great in big system
Delayed Correlation vs Causation

![Graph showing similarity between A and R as a function of network size](image)
Omitted Variable Bias

\[ y_i = x_i \beta + z_i \delta + u_i \]

\[ \hat{\beta} = (X' X)^{-1} X' (X \beta + Z \delta + U) \]

\[ E[\hat{\beta} \mid X] = \beta + (X' X)^{-1} E[X' Z \mid X] \delta \]

The bias should be arbitrarily big relative to the signal.
This problem does not go away with more data.

Measuring and interpreting neuronal correlations

Marlene R. Cohen\textsuperscript{1} and Adam Kohn\textsuperscript{2}
An ideal RCT allows unbiased measurements of causal effects
But alas, the world is not perfect: observational studies

Why is it so hard to obtain causality?
Instrumental variables

Quantifying causality in data science with quasi-experiments

Tony Liu,¹ Lyle Ungar,¹ and Konrad Kording²,³,₈
If exclusion restriction correct
If it is violated
There are lots of quasiexperiments

- Check out Tony’s tutorials

https://github.com/tliu526/causal-data-science-perspective
Does eating grains make ducks lay more eggs?
IV analysis strategy

Confounding

Instrument $Z$ → Treatment $T$ → Outcome $Y$
Problem: non-compliance

Never-takers

Always-takers

Compliers
Improving exclusion criteria for IV studies

\[ Var_{IV} = \frac{Var[Y|Z, compliers]}{N_{comply} p^2 E[Z](1-E[Z])} \]

\[ n = 10 \]
Improving exclusion criteria for IV studies

\[ n_{\text{eff}} = 3.6 \]

\[ N_{\text{eff}} = N \times p_{\text{comply}}^2 \]
Improving exclusion criteria for IV studies

\[ N_{\text{eff}} = N \times p_{\text{comply}}^2 \]

\[ n_{\text{eff}} = 6 \]
Data-driven exclusion criteria

Use observed features and machine learning to predict out-of-sample compliance
Data-driven exclusion improves power
Applying this to diabetes

- Optum data
- Above A1C=6.5 you should be diabetic
- But it is fuzzy (hence IV)
- Follow-up A1C should be better
Let us use regression discontinuity design.
0.065 ± 0.006

0.095 ± 0.007
0.065 ± 0.006

-0.55 ± 0.17

0.095 ± 0.007

-0.49 ± 0.13
A nonstandard logic

- ML tells us whom we are talking about
- Our predictions then only apply to these items
- We can then purposefully say nothing about the expected noncompliers
- (which seems unavoidably true)
Find all the RDDs

- Currently working to
  - Apply this idea to every threshold
  - Continuously sweep to find all thresholds
  - Optum and other datasets
A bitter lesson approach

Learning domain-specific causal discovery from time series

Xinyue Wang
Department of Bioengineering
University of Pennsylvania

Konrad Kording
Department of Bioengineering
University of Pennsylvania

Published in Transactions on Machine Learning Research (09/2023)

wsingue@seas.upenn.edu
koerding@gmail.com
Supervised learning

- Know inputs and true output
- Get good at mapping from input to output
- Causal inference is not like this!
Where do CI/CD algorithms come from?

Just like in ML
Causal inference is supervised learning, after all.
We need something to test this idea

- Something with known causality
- And nontrivial data
A causal discovery problem with the microprocessor

A. Causal pairs

B. Non-causal pairs

- Read through two channels. Predict if they are connected
Perturbations to define causality
Honest validation

- Use 1 half of the data to train a causality detector
- Use other half to test if it works
Standard Transformer ML system
Works pretty great
Also robust to noise
Generalizes well across games: train on Donkey Kong, test others
Focuses attention where it matters (transients)
Simulated fMRI

Table 3: AUPRC comparison on different simulations of NetSim (mean ± std).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Corr</th>
<th>DYNO</th>
<th>GC</th>
<th>MI</th>
<th>PCMCI+</th>
<th>LiNGAM</th>
<th>cMLP</th>
<th>cLSTM</th>
<th>eSRU</th>
<th>SRU</th>
<th>Transformer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sim1</td>
<td>0.47 ± 0.02</td>
<td>0.41 ± 0.08</td>
<td>0.40 ± 0.08</td>
<td>0.39 ± 0.08</td>
<td>0.39 ± 0.09</td>
<td>0.43 ± 0.15</td>
<td>0.42 ± 0.15</td>
<td>0.41 ± 0.14</td>
<td>0.40 ± 0.14</td>
<td>0.39 ± 0.14</td>
<td>0.97 ± 0.06</td>
</tr>
<tr>
<td>Sim2</td>
<td>0.45 ± 0.03</td>
<td>0.33 ± 0.12</td>
<td>0.32 ± 0.12</td>
<td>0.31 ± 0.11</td>
<td>0.29 ± 0.11</td>
<td>0.30 ± 0.11</td>
<td>0.29 ± 0.11</td>
<td>0.28 ± 0.11</td>
<td>0.27 ± 0.11</td>
<td>0.26 ± 0.10</td>
<td>0.94 ± 0.05</td>
</tr>
<tr>
<td>Sim3</td>
<td>0.44 ± 0.03</td>
<td>0.32 ± 0.13</td>
<td>0.29 ± 0.14</td>
<td>0.27 ± 0.12</td>
<td>0.26 ± 0.12</td>
<td>0.27 ± 0.13</td>
<td>0.26 ± 0.12</td>
<td>0.24 ± 0.12</td>
<td>0.23 ± 0.12</td>
<td>0.22 ± 0.12</td>
<td>0.89 ± 0.06</td>
</tr>
<tr>
<td>Sim4</td>
<td>0.41 ± 0.07</td>
<td>0.36 ± 0.08</td>
<td>0.35 ± 0.11</td>
<td>0.37 ± 0.10</td>
<td>0.36 ± 0.10</td>
<td>0.40 ± 0.14</td>
<td>0.40 ± 0.14</td>
<td>0.39 ± 0.14</td>
<td>0.39 ± 0.14</td>
<td>0.38 ± 0.13</td>
<td>0.85 ± 0.13</td>
</tr>
<tr>
<td>Sim5</td>
<td>0.48 ± 0.02</td>
<td>0.38 ± 0.10</td>
<td>0.39 ± 0.12</td>
<td>0.40 ± 0.11</td>
<td>0.40 ± 0.12</td>
<td>0.42 ± 0.16</td>
<td>0.42 ± 0.16</td>
<td>0.42 ± 0.15</td>
<td>0.42 ± 0.15</td>
<td>0.42 ± 0.15</td>
<td>0.96 ± 0.03</td>
</tr>
<tr>
<td>Sim6</td>
<td>0.28 ± 0.03</td>
<td>0.26 ± 0.04</td>
<td>0.26 ± 0.06</td>
<td>0.25 ± 0.07</td>
<td>0.25 ± 0.08</td>
<td>0.24 ± 0.08</td>
<td>0.24 ± 0.08</td>
<td>0.23 ± 0.08</td>
<td>0.71 ± 0.10</td>
<td>0.90 ± 0.06</td>
<td></td>
</tr>
<tr>
<td>Sim7</td>
<td>0.43 ± 0.02</td>
<td>0.36 ± 0.08</td>
<td>0.33 ± 0.11</td>
<td>0.31 ± 0.10</td>
<td>0.29 ± 0.11</td>
<td>0.30 ± 0.11</td>
<td>0.28 ± 0.11</td>
<td>0.26 ± 0.11</td>
<td>0.26 ± 0.11</td>
<td>0.76 ± 0.10</td>
<td></td>
</tr>
<tr>
<td>Sim8</td>
<td>0.48 ± 0.04</td>
<td>0.47 ± 0.05</td>
<td>0.48 ± 0.07</td>
<td>0.49 ± 0.10</td>
<td>0.47 ± 0.10</td>
<td>0.48 ± 0.10</td>
<td>0.47 ± 0.10</td>
<td>0.47 ± 0.10</td>
<td>0.47 ± 0.10</td>
<td>0.46 ± 0.11</td>
<td>0.72 ± 0.09</td>
</tr>
<tr>
<td>Sim9</td>
<td>0.48 ± 0.02</td>
<td>0.41 ± 0.08</td>
<td>0.41 ± 0.09</td>
<td>0.40 ± 0.08</td>
<td>0.38 ± 0.09</td>
<td>0.42 ± 0.13</td>
<td>0.41 ± 0.13</td>
<td>0.40 ± 0.13</td>
<td>0.39 ± 0.13</td>
<td>0.93 ± 0.08</td>
<td></td>
</tr>
<tr>
<td>Sim10</td>
<td>0.45 ± 0.03</td>
<td>0.38 ± 0.07</td>
<td>0.40 ± 0.09</td>
<td>0.41 ± 0.10</td>
<td>0.48 ± 0.21</td>
<td>0.47 ± 0.20</td>
<td>0.45 ± 0.20</td>
<td>0.44 ± 0.19</td>
<td>0.43 ± 0.19</td>
<td>0.92 ± 0.09</td>
<td></td>
</tr>
<tr>
<td>Sim11</td>
<td>0.48 ± 0.01</td>
<td>0.44 ± 0.05</td>
<td>0.45 ± 0.07</td>
<td>0.44 ± 0.06</td>
<td>0.44 ± 0.06</td>
<td>0.46 ± 0.10</td>
<td>0.46 ± 0.10</td>
<td>0.45 ± 0.10</td>
<td>0.45 ± 0.10</td>
<td>0.96 ± 0.03</td>
<td></td>
</tr>
<tr>
<td>Sim12</td>
<td>0.47 ± 0.01</td>
<td>0.39 ± 0.09</td>
<td>0.36 ± 0.10</td>
<td>0.36 ± 0.09</td>
<td>0.35 ± 0.10</td>
<td>0.42 ± 0.19</td>
<td>0.40 ± 0.19</td>
<td>0.37 ± 0.19</td>
<td>0.35 ± 0.19</td>
<td>0.98 ± 0.02</td>
<td></td>
</tr>
<tr>
<td>Sim13</td>
<td>0.48 ± 0.03</td>
<td>0.42 ± 0.07</td>
<td>0.42 ± 0.12</td>
<td>0.41 ± 0.10</td>
<td>0.40 ± 0.11</td>
<td>0.43 ± 0.16</td>
<td>0.42 ± 0.16</td>
<td>0.41 ± 0.16</td>
<td>0.40 ± 0.15</td>
<td>0.99 ± 0.02</td>
<td></td>
</tr>
<tr>
<td>Sim14</td>
<td>0.48 ± 0.03</td>
<td>0.42 ± 0.08</td>
<td>0.40 ± 0.08</td>
<td>0.38 ± 0.09</td>
<td>0.42 ± 0.15</td>
<td>0.41 ± 0.14</td>
<td>0.40 ± 0.14</td>
<td>0.39 ± 0.14</td>
<td>0.38 ± 0.13</td>
<td>0.95 ± 0.06</td>
<td></td>
</tr>
<tr>
<td>Sim15</td>
<td>0.41 ± 0.04</td>
<td>0.38 ± 0.06</td>
<td>0.38 ± 0.08</td>
<td>0.39 ± 0.07</td>
<td>0.37 ± 0.08</td>
<td>0.37 ± 0.09</td>
<td>0.35 ± 0.09</td>
<td>0.34 ± 0.09</td>
<td>0.34 ± 0.09</td>
<td>0.31 ± 0.11</td>
<td></td>
</tr>
<tr>
<td>Sim16</td>
<td>0.40 ± 0.04</td>
<td>0.35 ± 0.06</td>
<td>0.40 ± 0.12</td>
<td>0.39 ± 0.10</td>
<td>0.41 ± 0.14</td>
<td>0.47 ± 0.21</td>
<td>0.45 ± 0.20</td>
<td>0.43 ± 0.19</td>
<td>0.42 ± 0.19</td>
<td>0.41 ± 0.19</td>
<td>0.41 ± 0.05</td>
</tr>
<tr>
<td>Sim17</td>
<td>0.36 ± 0.06</td>
<td>0.31 ± 0.07</td>
<td>0.34 ± 0.10</td>
<td>0.35 ± 0.10</td>
<td>0.35 ± 0.11</td>
<td>0.35 ± 0.11</td>
<td>0.34 ± 0.11</td>
<td>0.34 ± 0.11</td>
<td>0.33 ± 0.11</td>
<td>0.37 ± 0.11</td>
<td></td>
</tr>
</tbody>
</table>

43
Gene networks

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Corr</th>
<th>DYNO</th>
<th>GC</th>
<th>MI</th>
<th>PCMCI+</th>
<th>cMLP</th>
<th>cLSTM</th>
<th>eSRU</th>
<th>SRU</th>
<th>Transformer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ecoli2</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
<td>0.01</td>
<td>0.01</td>
<td>0.04</td>
</tr>
<tr>
<td>Yeast2</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
<td>0.01</td>
<td>0.01</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.06</td>
</tr>
<tr>
<td>Yeast3</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.06</td>
<td><strong>0.07</strong></td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
</tr>
</tbody>
</table>

Table 5: AUPRC comparison on the Dream3
What does this mean

- **Cons**
  - Requires a lot of ground truth data
  - Impossible for humans to understand
  - Biased

- **Pros**
  - If we have enough ground truth it will beat all humans
  - Meaningful proofs (inherited from empirical risk minimization)
  - A new approach to get into messy observational spaces
Take home message

- Causality and ML really have overlapping problems
- So much to learn
- Check out Clear conference

- Exclusion can be a meaningful data problem
- Maybe we should look for more ways to learn how to do CD/ CI, and prepare ourselves for a future where the bitter lesson is key to medicine