A nationally-implemented AI solution for Covid-19

Mihaela van der Schaar
John Humphrey Plummer Professor of Machine Learning, Artificial Intelligence and Medicine, University of Cambridge
Chancellor’s Professor, University of California Los Angeles

Email: mv472@cam.ac.uk
Website: http://www.vanderschaar-lab.com/
Agenda

Using AI/ML to fight Covid-19 at:
• Global Level - Policy Impact Predictor
• National Level - Resource Planning
• Patient Level - Interpretable and explainable AI/ML

Joining forces to transform healthcare!
Why is AI/ML for healthcare is different?

AI/ML has accomplished wonders .... on well-posed problems where the notion of a “solution” is well-defined and solutions are verifiable

Healthcare is different – problems are not well-posed and notion of a “solution” is often not well-defined and solutions are hard to verify

This presents enormous challenges – and also enormous opportunities

Goal: Augment human decision making
- clinicians, medical researchers and policy makers

New ML models and techniques
Covid-19 at the Global Level

Policy Impact Predictor

The problem:
Estimating the *causal effects of NPIs* applied over time on COVID-19 deaths to conduct *counterfactual scenario analysis*.

Why is it important?
*Inform governments* and policy-makers on what NPIs to apply over the next months.

What is new?
Learning *heterogeneous* NPI effects using *global* data from different countries.

Potential impact...
*Reduction of future COVID-19 deaths* in various countries around the world.
### Current Modeling Efforts: CDC National Forecasts*

<table>
<thead>
<tr>
<th></th>
<th>Model Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>IHME</strong></td>
<td><em>University of Washington</em>)&lt;br&gt;Combination of a mechanistic disease transmission model and a curve-fitting approach.</td>
</tr>
<tr>
<td><strong>Imperial</strong></td>
<td><em>Imperial College London</em>&lt;br&gt;Ensembles of mechanistic transmission models, fit to different parameter assumptions.</td>
</tr>
<tr>
<td><strong>LANL</strong></td>
<td><em>Los Alamos National Laboratory</em>&lt;br&gt;Statistical dynamical growth model accounting for population susceptibility.</td>
</tr>
<tr>
<td><strong>DELPHI</strong></td>
<td><em>Massachusetts Institute of Technology</em>&lt;br&gt;SEIR model.</td>
</tr>
<tr>
<td><strong>YYG</strong></td>
<td><em>Independent</em>&lt;br&gt;SEIS mechanistic model</td>
</tr>
<tr>
<td><strong>CDC-ensemble</strong></td>
<td><em>Centre for Disease Control and Prevention</em>&lt;br&gt;An ensemble of 21 individual forecasts.</td>
</tr>
</tbody>
</table>

Problem Formulation

- COVID-19 deaths over time: \( Y_i(t) \in \mathbb{N} \cup \{0\} \) for country \( i \in \{1, \ldots, N\} \)

- Country-level features \( X_i(t) \in \mathbb{R}^d \)
  - Country meta-data \( X_i^s \)
  - Exogenous time-varying features \( X_i^{ex}(t) \)
  - Endogenous time-varying features \( X_i^{en}(t) \)

- NPI indicators: \( \mathcal{P}_i[1 : t] \triangleq \{p_i(1), \ldots, p_i(t)\} \)
  - \( p_i(t) \in \{0, 1\}^K \)
  - \( K \) policy indicators

Dataset
\[ \mathcal{D}_{N,t} \triangleq \{X_i, Y_i[1 : t], \mathcal{P}_i[1 : t]\}_{i=1}^N \]

Demographic, social, economic, environmental, and public health indicators
Weather patterns
Community mobility levels
Problem Formulation

- **Potential outcomes framework**

  - Interventions: \( p_i(t) = [p_i^1(t), \ldots, p_i^K(t)] \)
  - Potential outcomes: \( \{\mathcal{Y}_i[p^k=1][t : t + T], \mathcal{Y}_i[p^k=0][t : t + T]\} \)
  - Confounders: \( \mathcal{H}_{t-1} \triangleq \{X_i, \mathcal{Y}_i[1 : t - 1], P_i[1 : t - 1]\} \)

- **Assumptions**

  - No unobserved confounders

\[ \mathcal{Y}_i[p^k][t : t + T] \perp p_i(t) \mid \mathcal{H}_{t-1} \]
COVID-19 Counterfactual Scenario Analysis

For a given country $i \in \{1, \ldots, N\}$, forecast the trajectory of COVID-19 deaths within a future time horizon of $T$ days under a given set of future NPIs

$$\hat{Y}_i[t : t + T] = \mathbb{E} \left[ Y_i[t : t + T] \mid p_i(t), \mathcal{H}_{t-1} \right]$$

Timeline

Policy-making window

Policy decision

$\hat{Y}_i[t : t + T]$

March April May June July August

- Public events cancellation
- Public transport closure
- International travel restriction
- Stay-at-home orders
Global Modeling: Exploit Heterogeneity of NPIs

Introduction of NPIs over time in 5 countries with most COVID-19 cases

Different types of NPIs
- Public events cancellation
- Public transport closure
- Debt relief
- Gatherings restriction
- Income support
- Internal movement restrict.
- International travel restrict.
- Public information camp.
- School closure
- Stay-at-home orders
- Testing policy
- Workplace closure
- Contact tracing
- Masks

Different timing for NPIs

Brazil
- Days since first case
- Public events cancellation
- Public transport closure
- Debt relief
- Gatherings restriction
- Income support
- Internal movement restrict.
- International travel restrict.
- Public information camp.
- School closure
- Stay-at-home orders
- Testing policy
- Workplace closure
- Contact tracing
- Masks

Russia
- Days since first case
- Public events cancellation
- Public transport closure
- Debt relief
- Gatherings restriction
- Income support
- Internal movement restrict.
- International travel restrict.
- Public information camp.
- School closure
- Stay-at-home orders
- Testing policy
- Workplace closure
- Contact tracing
- Masks

Spain
- Days since first case
- Public events cancellation
- Public transport closure
- Debt relief
- Gatherings restriction
- Income support
- Internal movement restrict.
- International travel restrict.
- Public information camp.
- School closure
- Stay-at-home orders
- Testing policy
- Workplace closure
- Contact tracing
- Masks

United Kingdom
- Days since first case
- Public events cancellation
- Public transport closure
- Debt relief
- Gatherings restriction
- Income support
- Internal movement restrict.
- International travel restrict.
- Public information camp.
- School closure
- Stay-at-home orders
- Testing policy
- Workplace closure
- Contact tracing
- Masks

United States
- Days since first case
- Public events cancellation
- Public transport closure
- Debt relief
- Gatherings restriction
- Income support
- Internal movement restrict.
- International travel restrict.
- Public information camp.
- School closure
- Stay-at-home orders
- Testing policy
- Workplace closure
- Contact tracing
- Masks

Days since first case
Modeling COVID-19 Death Curves

- Model the trajectory of COVID-19 deaths within country $i$ as a **Gaussian process**

  \[ f_{L,i} \sim \mathcal{GP}(D_{\theta_i}(t), K_{\theta_i}(t, t')) \]

  - Mean function: compartmental model
    \[ D_{\theta_i}(t) \]
    - SEIR model
  - Kernel function
    \[ K_{\theta_i}(t, t') = \exp\left(\frac{-1}{2} \| t - t' \|^2 \right) \]

- Daily COVID-19 deaths in China
- Posterior mean and variance
- 100 deaths
Modeling COVID-19 Death Curves

- Model the trajectory of COVID-19 deaths within country $i$ as a Gaussian process

  \[ f_{L,i} \sim \mathcal{GP}(D_{\theta_i}(t), K_{\theta_i}(t, t')) \]

- Mean function = compartmental model

  \[ D_{\theta_i}(t) \]  
  \[ \text{SEIR model} \]

- Death forecasts = posterior trajectory of deaths.
- Prior = rigorous mechanistic model.
- Posterior = data-driven.
- Posterior variance = Uncertainty in projections.

Daily COVID-19 deaths in China
Compartmental Prior

- **Mean function =** 
  Susceptible, Exposed, Infectious, Recovered (SEIR) model

\[
f_{L,i} \sim \mathcal{GP}(D_{\theta_i}(t), K_{\theta_i}(t,t')),
\]

\[
\frac{dS_i}{dt} = \mu_i (n_i - S_i) - \frac{\beta_i S_i I_i}{n_i},
\]

\[
\frac{dE_i}{dt} = \frac{\beta_i S_i I_i}{n_i} - (\mu_i + \sigma) E_i,
\]

\[
\frac{dI_i}{dt} = \sigma E_i - (\gamma_i + \mu_i) I_i
\]

\[
\frac{dD_i}{dt} = \mu_i I_i
\]

- Contact rate
- Deceased

\[
\beta_i(t)
\]

\[
D_i(t)
\]

\[
\mu_i
\]

\[
\gamma_i
\]

**Diagram:**

- **Susceptible:** $S_i(t)$
- **Exposed:** $E_i(t)$
- **Infected:** $I_i(t)$
- **Recovered:** $R_i(t)$
Modeling Potential Outcomes

- Reproduction number: $R_{0,i}(t) = \frac{\sigma}{\mu_i + \sigma} \cdot \frac{\beta_i(t)}{\mu_i + \gamma_i}$

- NPI effects on potential outcomes = “Flattening” the curve.

![Diagram of susceptible, infected, recovered, and deceased flows with contact rate and reproduction number equations]
Joint Model for all Countries

- Hierarchical Bayesian model
  - Multi-layered model

**Upper-layer Gaussian process**

\[ f_U \sim GP(m_\alpha(X, p), K_\alpha((X, p), (X', p'))) \]

NPI variables and country-level features, shared among all countries

**Lower-layer Gaussian process**

\[ f_{L,i} \sim GP(D_{\theta_i}(t), K_{\theta_i}(t, t')), \]
\[ \theta_i = v(f_U(X_i, p_i)) \]

Specific to a country, models deaths over time
Model: Impact of Non-pharmaceutical Interventions

- Reproduction number: \( R_{0,i}(t) = \frac{\sigma}{\mu_i+\sigma} \cdot \frac{\beta_i(t)}{\mu_i+\gamma_i} \)  
  Baseline \( R_0 \) depends on country’s features

- NPIs effects depend on the country’s features...

\[
\beta_i(t) = \nu(f_U(X_i, p_i(t))) = 2\tilde{\beta}\text{Sigmoid}(f_U(X_i, p_i(t)))
\]  
  Contact rate

<table>
<thead>
<tr>
<th>Country feature</th>
<th>Loose NPIs</th>
<th>Strict NPIs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brazil</td>
<td><img src="image" alt="Brazil" /></td>
<td></td>
</tr>
<tr>
<td>United States</td>
<td><img src="image" alt="United States" /></td>
<td><img src="image" alt="United States" /></td>
</tr>
</tbody>
</table>

Contact rate

Reproduction number \( R_0 \)

NPI

Country feature \( X \)
A Dataset for Global COVID-19 Trajectories

- Country-level COVID-19 data from multiple sources was collated for **170 countries**.

  - COVID-19 cases, deaths, tests, non-pharmaceutical interventions, excess mortality, mobility statistics, weather patterns, country meta-data.

<table>
<thead>
<tr>
<th>Category</th>
<th>Sources/Platforms</th>
</tr>
</thead>
<tbody>
<tr>
<td>COVID-19 cases, deaths and tests</td>
<td>European center for disease prevention and control (ECDC)</td>
</tr>
<tr>
<td>Non-pharmaceutical interventions</td>
<td>Oxford Blavatnik school of government, ACAPS Gov. measures</td>
</tr>
<tr>
<td>Excess mortality</td>
<td>Human mortality Database, EuroStat, The Economist</td>
</tr>
<tr>
<td>Mobility statistics</td>
<td>Google community mobility reports, Apple maps</td>
</tr>
<tr>
<td>Weather patterns</td>
<td>UK Met office</td>
</tr>
<tr>
<td>Country-level Meta-data</td>
<td>World Bank reports</td>
</tr>
</tbody>
</table>
Non-pharmaceutical interventions (NPI)

- The Oxford Government Response Tracker provides 13 NPIs on an ordinal scale for each country, reflecting both the intensity of enforcement of the intervention.

<table>
<thead>
<tr>
<th>Public events cancellation</th>
<th>Internal movement restrictions</th>
<th>Stay-at-home orders</th>
</tr>
</thead>
<tbody>
<tr>
<td>Public transport closure</td>
<td>International travel restrictions</td>
<td>Testing policy</td>
</tr>
<tr>
<td>Gatherings restrictions</td>
<td>Public information campaigns</td>
<td>Workplace closure</td>
</tr>
<tr>
<td>Income support</td>
<td>School closure</td>
<td>Contact tracing</td>
</tr>
</tbody>
</table>

Mandating on mask usage
## Country-level Meta-data

- **We collected 35 economic, social, demographic, environmental and public health indicators from published World Bank reports**

<table>
<thead>
<tr>
<th>Category</th>
<th>Indicators</th>
</tr>
</thead>
<tbody>
<tr>
<td>Economic</td>
<td>GDP per capita, GNI per capita, Income share held by lowest 20%</td>
</tr>
<tr>
<td>Social and demographic</td>
<td>Population, Life expectancy, Birth rate, Death rate, Infant mortality rate, Land Area, % People with basic hand-washing facilities including soap and water, Smoking prevalence, Prevalence of undernourishment, Prevalence of overweight, Urban population, Population density, Population ages 65 and above, Access to electricity (% of population), UHC service coverage index, Total alcohol consumption per capita, Air transport (passengers carried)</td>
</tr>
<tr>
<td>Environmental</td>
<td>Forest Area, PM2.5 air pollution (mean annual exposure in micrograms per cubic meter)</td>
</tr>
<tr>
<td>Public health</td>
<td>Immunization for measles, % deaths by communicable diseases, Current health expenditure, Current health expenditure per capita, Diabetes prevalence, Immunization for DPT, Immunization for HepB3, Incidence of HIV, Incidence of malaria, Incidence of tuberculosis, % deaths by CVD/cancer/diabetes/CRD, % deaths due to household and ambient air pollution, % deaths due to unsafe water/unsafe sanitation/lack of hygiene, Physicians (per 1,000 people)</td>
</tr>
</tbody>
</table>
## Results: Accuracy of US Projections

<table>
<thead>
<tr>
<th>Model</th>
<th>March 28 forecasts (before the peak)</th>
<th>April 11 forecasts (During the peak)</th>
<th>April 25 forecasts (After the peak)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>7 days</td>
<td>14 days</td>
<td>7 days</td>
</tr>
<tr>
<td>YYG</td>
<td>___</td>
<td>___</td>
<td>-6,470</td>
</tr>
<tr>
<td>Imperial</td>
<td>___</td>
<td>___</td>
<td>1,757</td>
</tr>
<tr>
<td>LANL</td>
<td>___</td>
<td>___</td>
<td>-6,010</td>
</tr>
<tr>
<td>MIT-DELPHI</td>
<td>___</td>
<td>___</td>
<td>2,174</td>
</tr>
<tr>
<td>Gompertz curve</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>Vanilla SEIR</td>
<td>2,723</td>
<td>4,822</td>
<td>-11,328</td>
</tr>
<tr>
<td>IHME</td>
<td>-1,999</td>
<td>-2,289</td>
<td>-6,134</td>
</tr>
<tr>
<td>CDC-ensemble</td>
<td>___</td>
<td>___</td>
<td>-2,739</td>
</tr>
<tr>
<td>PIP model (US only)</td>
<td>-642</td>
<td>-4,380</td>
<td>-3,182</td>
</tr>
<tr>
<td>PIP model (global)</td>
<td>-867</td>
<td>-1,396</td>
<td>-1,906</td>
</tr>
</tbody>
</table>
## Results: Accuracy of Global Projections

<table>
<thead>
<tr>
<th></th>
<th>March 28 forecasts</th>
<th></th>
<th>April 11 forecasts</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IHME</td>
<td>YYG</td>
<td>Imperial</td>
<td>PIP model</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>-981</td>
<td>-3,479</td>
<td>-182</td>
<td><strong>-131</strong></td>
</tr>
<tr>
<td>Italy</td>
<td>-1,082</td>
<td>451</td>
<td>1,804</td>
<td><strong>294</strong></td>
</tr>
<tr>
<td>Germany</td>
<td>-420</td>
<td>244</td>
<td>-417</td>
<td><strong>104</strong></td>
</tr>
<tr>
<td>Spain</td>
<td>1,104</td>
<td><strong>167</strong></td>
<td>-499</td>
<td>317</td>
</tr>
<tr>
<td>Brazil</td>
<td>___</td>
<td>-283</td>
<td>___</td>
<td><strong>-105</strong></td>
</tr>
<tr>
<td>Sweden</td>
<td>311</td>
<td>107</td>
<td>-102</td>
<td><strong>24</strong></td>
</tr>
<tr>
<td>France</td>
<td>-501</td>
<td>803</td>
<td>-2,415</td>
<td><strong>-79</strong></td>
</tr>
<tr>
<td>Netherlands</td>
<td>512</td>
<td>172</td>
<td>265</td>
<td><strong>-21</strong></td>
</tr>
<tr>
<td>Iran</td>
<td>___</td>
<td>40</td>
<td>___</td>
<td><strong>9</strong></td>
</tr>
<tr>
<td>Mexico</td>
<td>___</td>
<td>-82</td>
<td>___</td>
<td><strong>-56</strong></td>
</tr>
<tr>
<td>Japan</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td><strong>-3</strong></td>
</tr>
<tr>
<td>South Africa</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td><strong>-8</strong></td>
</tr>
</tbody>
</table>
Counterfactual NPI Scenario Analysis using PIP

What could have happened?

What could happen?

Covid-19 Deaths

Timeline

Strict policy
Loose policy
Actual policy

Policy-making window
News

Trials begin of machine learning system to help hospitals plan and manage COVID-19 treatment resources developed by NHS Digital and University of Cambridge

Trials have begun of a system that will use machine learning to help predict the upcoming demand for intensive care (ICU) beds and ventilators needed to treat patients with COVID-19 at individual hospitals and across regions in England.

Date: 20 April 2020
Adjutorium: AutoPrognosis for Covid-19

Our goal: Provide evidence that reliably assists the difficult decisions clinicians and managers have to make to save lives

Use depersonalized data
• demographic info, comorbidities, hospitalization details, outcomes

to:
• forecast personalized risk for each patient
• forecast personalized patient benefit from resources
• forecast which treatments are needed by each patient and when
• forecast which resources are needed by each patient and when
• forecast future resource requirements at the hospital level

www.vanderschaar-lab.com/covid-19/
Decisions that healthcare professionals need to make

- Which patients should be on ventilators, and for how long?
- Which patients should go to the ICU?
- When will patients be discharged?
- Which patients can safely go home?
How?

Many diseases, many variables, various needs! All is changing!

Can’t craft a model for each disease!

Make Machine Learning DO the Crafting

- Previous AutoML? Auto-WEKA and Auto-Sklearn
  - Limited performance gains
  - Meta-learning
  - Simplistic handling of missing data
  - Do not capture uncertainty
  - Limited to classification problems (survival, competing risks, time-series etc.)
AutoPrognosis [Alaa & vdS, ICML 2018]: A tool for crafting Clinical Scores

We need an entire pipeline!
Automated ML for clinical analytics

AutoPrognosis
ICML 2018
Cystic Fibrosis - Scientific Reports - 2018
UK Biobank - Plos One 2018
Breast Cancer – 2019
Covid - 2020

AutoPrognosis
Lee, Alaa, Zame, vdS, AISTATS 2019
Alaa, vdS, NIPS 2017
Lee, Yoon, vdS, TBME 2019
Alaa, vdS, ICML 2019
Zhang, Jarett, vdS, AISTATS 2020

Clairvoyance [Jarrett, Yoon, Bica, Ercole, vdS, 2020]
AutoPrognosis at work
Covid-19: Who needs ventilation?

**AUC-ROC accuracy for predicting whether a patient will need ventilation based on info available at hospital admission**

<table>
<thead>
<tr>
<th>Model</th>
<th>AUC-ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AP</strong>: all features</td>
<td>0.771 ± 0.002</td>
</tr>
<tr>
<td><strong>AP</strong>: age + specific comorbidities</td>
<td>0.761 ± 0.001</td>
</tr>
<tr>
<td><strong>AP</strong>: age + no. of comorbidities</td>
<td>0.720 ± 0.003</td>
</tr>
<tr>
<td><strong>Cox Regression</strong>: all features</td>
<td>0.690 ± 0.002</td>
</tr>
<tr>
<td><strong>Charlson Comorbidity Index</strong></td>
<td>0.618 ± 0.002</td>
</tr>
</tbody>
</table>
Démonstrateur
Covid-19 at the Patient Level

Prediction Results

Adjutorium is a system designed to assist decision-making by offering predictions based on existing data. It can support and inform healthcare professionals, but is not intended to replace their own decision-making processes.

Ventilator benefit
This table shows the expected survival chance for 14 days after diagnosis based on the patient information.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Additional Benefit</th>
<th>Overall Survival %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Invasive Ventilator</td>
<td>20%</td>
<td>84%</td>
</tr>
<tr>
<td>Invasive Ventilator</td>
<td>25%</td>
<td>87%</td>
</tr>
</tbody>
</table>

Build the ecosystem!
How can we turn ML models into actionable intelligence?

We need

- **Risk score understanding**: Users need to understand, quantify and manage risk
- **Transparency**: Users need to comprehend how the model makes predictions
- **Avoid implicit bias**: Users need to be able to check whether the model does not learn biases
- **Discovery**: Users need to distil insights and new knowledge from the learned model
- **Know what we do not know**: Users need to have a quantification of the model’s prediction uncertainty
Many kinds of interpretations exist...

Current methods are tailored to *one* type of interpretation

Uncovering one of the following

- What features are globally important, i.e. for the entire population?
- What features are locally important, i.e. for this patient?
- Feature interaction
- Model non-linearity

**Desiderata**

- Model-independent: general, not tailored to specific models
- Post-hoc: should not interfere with model training which may introduce bias and compromise accuracy
Which features of an individual are relevant for a prediction?

Mary

Age, Gender, Diabetes, Hypertension, SBP, ....

Black Box Predictive Model

Mortality due to Covid-19: 0.78

INVASE

Age, Diabetes, Hypertension, SBP

[Yoon, Jordon, vdS, ICLR 2019]
How can we learn individualized feature importance?

Key idea: Use Reinforcement Learning (RL)
- Make observations
- Select “actions” on the basis of these observations
- Determine “rewards” for these actions

Ultimately learn a policy which selects the best actions
  - i.e. actions that maximize rewards given observations

We use the Actor-Critic approach to RL
• **Selector network (actor)** takes instances and outputs vector of selection probabilities.
• **Predictor network (critic)** receives the selected features, makes predictions and provides feedback to the actor.
Find selector function $S$ that minimizes features selected $S(x)$ while satisfying equality constraints on the conditional distribution of the predictions.

- **Objective:** minimize $S(x)$
- **Constraints:**
  \[
  (Y|X(S(x)) = x(S(x))) \overset{d}{=} (Y|X = x)
  \]

- $x$: Features for a given realization
- $S: \mathcal{X} \to \{0,1\}^d$: Selector function, $S(x)$: Selected features
- $Y$: Predictions made by black-box model
INVASE: Instance-wise feature importance for prediction

Find selector function $S$ that minimizes features selected $S(x)$ while satisfying equality constraints on the conditional distribution of the predictions.

- **Objective**: minimize $S(x)$
- **Constraints**:
  \[ (Y|X^{(S(x))} = x^{(S(x))}) \overset{d}{=} (Y|X = x) \]

- **Lagrangian optimization**:
  \[ \mathcal{L}(S) = \mathbb{E}[KL(Y|X^{(S(x))} = x^{(S(x))})] + \lambda ||S(x)|| \]

- **Challenging problem**:
  - Output space of the selector function is large - its size increases exponentially with the dimension of the feature space!
  - We do not have access to the densities required – need to be learned [Yoon, Jordon, vdS, ICLR 2019]
## Limitations of past methods for model interpretability

<table>
<thead>
<tr>
<th>Method</th>
<th>Feature importance</th>
<th>Individualized feature importance</th>
<th>Model-independent</th>
<th>Identifying the set of relevant features for each instance</th>
</tr>
</thead>
<tbody>
<tr>
<td>LASSO</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>[Tibshirani, 1996]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knock-off</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>[Candes et al, 2016]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L2X</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>[Chen et al, 2018]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIME</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>[Ribeiro et al, 2016]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SHAPE</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>[Lundberg et al, 2017]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DeepLIFT</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>[Shrikumar et al, 2017]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Saliency</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>[Simonyan et al, 2013]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TreeSHAP</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>[Lundberg et al, 2018]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pixel-wise</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>[Batch et al, 2015]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INVASE</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>[Yoon, Jordon and van der Schaar, 2019]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

INVASE discovers the number of relevant features for each instance.
Are we done?

• **NO!**

• Need to ALSO understand what the model discovered: feature/statistical interactions, model non-linearity, etc.

<table>
<thead>
<tr>
<th>Method</th>
<th>Feature importance</th>
<th>Individualized feature importance</th>
<th>Feature interaction</th>
<th>Model-independent</th>
<th>Post-hoc</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIME</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>[Ribeiro et al, 2016]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SHAPE</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>[Lundberg et al, 2017]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DeepLIFT</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[Shrikumar et al, 2017]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INVASE</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>[Yoon, Jordon and van der Schaar, 2019]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L2X</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[Chen et al, 2018]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAM</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[Lou et al, 2013]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NIT</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[Tsang et al, 2018]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
What we are aiming for?

- Understand what the model discovered: feature importance, instance-wise feature importance, feature/statistical interactions, model non-linearity, etc.

- Produce a transparent risk equation describing the model for approval in practice guidelines

- Enable model explainability, not only interpretability

Can we have it all??

YES!
Demystifying Black-box Models with Symbolic Metamodels
[A. Alaa & vdS, NeurIPS 2019]

- **Metamodel** = a model of a model.
- **A symbolic metamodel outputs a transparent function describing the predictions of the black box model.**
- **Metamodelling needs only query access to trained black-box model.**
Symbolic Metamodelling

Model space

Metamodelling space

White-box model

Metamodelling space

Black-box ML model

Model space (uninterpretable)

Metamodelling space can be chosen by the user!
How are we going to achieve this?

- **Kolmogorov-Arnold Theorem** [Kolmogorov et al, 1961]

  Every multivariate continuous function can be written as a finite composition of univariate continuous functions

  \[
  g(x) = \sum_{q=0}^{r} g_q \left( \sum_{p=1}^{n} g_{q,p}(x_p) \right)
  \]

- **The symbolic metamodelling problem**

  **Metamodel representation**

  \[
  g(x; \theta) = \sum_{q=0}^{2n} G \left( \sum_{p=1}^{n} G(x_p; \theta_{q,p}); \theta_q \right)
  \]

  **Metamodel optimization**

  \[
  \theta^* = \arg \min_{\theta \in \Theta} \ell(f(x), g(x; \theta))
  \]
What basic functions?

- **Meijer G-functions** [C. S. Meijer, 1936]

\[
G_{p,q}^{m,n} \left( \begin{array}{c}
\frac{a_1, \ldots, a_p}{b_1, \ldots, b_q} \\
\end{array} | x \right) = \frac{1}{2\pi i} \int_L \frac{\prod_{j=1}^{m} \Gamma(b_j - s) \prod_{j=1}^{n} \Gamma(1 - a_j + s)}{\prod_{j=m+1}^{m+q} \Gamma(1 - b_j + s) \prod_{j=n+1}^{n+q} \Gamma(a_j - s)} \ x^s \ ds
\]

- Very general class of functions

- Parameter selection yields many familiar functions

<table>
<thead>
<tr>
<th>G-function</th>
<th>Equivalent function</th>
<th>G-function</th>
<th>Equivalent function</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G_{0,0}^{1,0} \left( 0 \bigg</td>
<td>-x \right)$</td>
<td>$e^x$</td>
<td>$G_{2,2}^{1,2} \left( \frac{1}{2}, \frac{1}{2}, 0 \bigg</td>
</tr>
<tr>
<td>$G_{2,2}^{1,2} \left( \frac{1}{2}, \frac{1}{2}, 0 \bigg</td>
<td>x \right)$</td>
<td>$\log(1 + x)$</td>
<td>$G_{1,2}^{2,0} \left( \frac{1}{2}, \frac{1}{2}, 0 \bigg</td>
</tr>
<tr>
<td>$G_{0,2}^{1,0} \left( 0, \frac{1}{2} \bigg</td>
<td>\frac{x^2}{4} \right)$</td>
<td>$\frac{1}{\sqrt{\pi}} \cos(x)$</td>
<td>$G_{1,2}^{2,0} \left( 0, \frac{1}{2} \bigg</td>
</tr>
<tr>
<td>$G_{0,2}^{1,0} \left( \frac{1}{2}, 0 \bigg</td>
<td>\frac{x^2}{4} \right)$</td>
<td>$\frac{1}{\sqrt{\pi}} \sin(x)$</td>
<td>$G_{0,2}^{1,0} \left( \frac{1}{2}, \frac{1}{2} \bigg</td>
</tr>
</tbody>
</table>
Building a symbolic metamodel

Metamodel construction is “analogous” to a 2-layer neural network

Parameters of a Meijer-G function can be learned by gradient descent! This can be done very fast!
Interpretability using symbolic metamodeling in practice

<table>
<thead>
<tr>
<th>Method</th>
<th>AUC-ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cox</td>
<td>0.690 ± 0.002</td>
</tr>
<tr>
<td>AutoPrognosis</td>
<td>0.771 ± 0.002</td>
</tr>
</tbody>
</table>
Example: Use Metamodules for Individual-level feature importance

Original Metamodle

\[ g(x) = \alpha_0 \text{Age} + \alpha_1 \text{BMI}^2 + \alpha_2 \text{Age} \cdot \text{BMI} + \alpha_3 \text{Age} \cdot \text{Gender} + \alpha_4 \text{Gender} \cdot (1 + \alpha_5 \text{Diabetes}) + \alpha_6 \log(\text{Age} \cdot \text{Diabetes} + 1) \]

Individual-level feature importance

\[ \frac{\partial g(x)}{\partial \text{Age}} = \alpha_0 + \alpha_2 \text{BMI} + \alpha_3 \text{Gender} + \frac{\alpha_6 \text{Diabetes}}{\text{Age}+1} \]

Beyond current feature importance
Explainability = User-dependent Interpretability

- Different users seek different forms of “understanding”…

- **Clinician**
  - Why a treatment is recommended for the patient at hand?

- **Researcher**
  - Data-Induced hypothesis

- **Patient**
  - Informed Consent
  - Life-style Changes
Metamodels: How to use them?

Different forms of interpretations can be extracted from a Metamodel’s forward and backward views!

- **Forward use**
  - Input = features
  - Output = risk
  - Treatment justification
  - Hypothesis induction
  - Variable interactions
  - Variable importance

- **Backward use**
  - Input = reduced risk
  - Output = features
  - Modifiable variables
  - Dosage recommendation
  - Policy design

\[
g(Age, ER, HER2, Tumor size, Nodes) \]
Regardless of the model $f(x)$, $g(x)$ is always a symbolic expression.

Unified format for many different types of black-box models: identify common discoveries by comparing their Metamodells.

Neural network

Random forest

SVM

Report common discoveries

$\alpha_1 X_1 + \alpha_2 X_2^2 + \alpha_3 X_1 X_2$

$\alpha_4 X_3^3 + \alpha_5 \log(X_4)$

$\alpha_1 X_1 + \alpha_2 X_2^2 + \alpha_3 X_1 X_2$

$\alpha_4 X_3^3 + \alpha_5 \log(X_4)$

$\alpha_1 X_1 + \alpha_2 X_2^2 + \alpha_3 X_1 X_2$

$\alpha_4 X_3^3 + \alpha_5 \log(X_4)$
How can clinicians use Metamodels?

- Understand why how predictions or treatment recommendations are being made by the ML-model
- **Example:** Two patients with apparently similar features get different treatment recommendations!

Explanation:
Small changes in number of Lymph node causes a quadratic increase in risk
How can patients use Metamodels?

 Patients can be informed how to alter behavior to lower risk.

Can be set through the inverse Metamodel equation

\[
\text{BMI Reduction} = g^{-1}(\text{Family history, Genetics, Diabetes}| \text{Risk} = X \%)
\]
In addition to interpretability & explainability…. trustworthiness is key
Our approach: Post-hoc methodology with frequentist coverage guarantees

<table>
<thead>
<tr>
<th>Method</th>
<th>Post-hoc vs Built-in</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayesian neural nets (Ritter et al., 2018)</td>
<td>Built-in</td>
<td>No guarantees</td>
</tr>
<tr>
<td>Probabilistic backprop. (Blundell et al., 2015)</td>
<td>Built-in</td>
<td>No guarantees</td>
</tr>
<tr>
<td>Monte Carlo dropout (Gal &amp; Ghahramani, 2016)</td>
<td>Built-in</td>
<td>No guarantees</td>
</tr>
<tr>
<td>Deep Ensembles (Lakshminarayanan et al., 2017)</td>
<td>Built-in</td>
<td>No guarantees</td>
</tr>
<tr>
<td><strong>Discriminative Jackknife</strong> (Alaa and vdS, ICML2020)</td>
<td>Post-hoc 1-α</td>
<td></td>
</tr>
</tbody>
</table>

Does not interfere with model training or compromise accuracy!

*Alaa and vdS, ICML2020 - Frequentist Uncertainty in Recurrent Neural Networks via Blockwise Influence Functions*
Machine Learning & Healthcare: Vision

Clinical Practice

Linked EHR Data

Clinical Research

Observational Data

Actionable Intelligence

Data-induced Causal Discovery

Data-induced Genetic associations

Data-induced Hypotheses

Machine Learning

Augmented MD

Pharma

OMICS
Transforming healthcare using ML

- New understanding of diseases (Think Covid-19!)
- New understanding of relationships among diseases (multiple morbidities)
- New understanding of effects of interventions/treatments
- New ways of screening and monitoring
- New ways of preventing disease
- New ways of diagnosing and staging disease
- New ways of treating disease
  - Impact on clinical trials
- New ways of allocating resources
- New pathways of care

Join us!
Acknowledgements

Dr. Ahmed Alaa  Ioana Bica
Dr. Hyunsuk Lee  Zhaozhi Qian
Jinsung Yoon  Trent Kyono
Alexis Bellot  James Jordon
Changhee Lee  Dan Jarrett
Yao Zhang  Alihan Huyuk

Email: mv472@cam.ac.uk
Website: http://www.vanderschaar-lab.com/
We are hiring brilliant students and post-docs