Friends Don’t Let Friends Deploy Black-Box Models: Detecting and Preventing Bias via Transparent Modeling

Rich Caruana
Microsoft Research

Joint Work with
Yin Lou & Sarah Tan
Johannes Gehrke, Paul Koch, Marc Sturm, Noemie Elhadad

Thanks to
Greg Cooper MD PhD, Mike Fine MD MPH, Eric Horvitz MD PhD
Nick Craswell, Tom Mitchell, Jacob Bien, Giles Hooker, Noah Snavely
When is it Safe to Use Machine Learning in Healthcare?

- data for 1M patients
- 1000’s great clinical features
- train state-of-the-art machine learning model on data
- accuracy looks great on test set: AUC = 0.95

- is it safe to deploy this model and use on real patients?
- is high accuracy on test data enough to trust a model?
When is it Safe to Use Machine Learning in Healthcare?

- data for 1M patients
- 1000’s great clinical features
- train state-of-the-art machine learning model on data
- accuracy looks great on test set: AUC = 0.95

is it safe to deploy this model and use on real patients?

is high accuracy on test data enough to trust a model?
When is it Safe to Use Machine Learning in Healthcare?

- data for 1M patients
- 1000’s great clinical features
- train state-of-the-art machine learning model on data
- accuracy looks great on test set: AUC = 0.95

- is it safe to deploy this model and use on real patients?
- is high accuracy on test data enough to trust a model?
Motivation: Predicting Pneumonia Risk Study (mid-90’s)

- **LOW Risk:** outpatient: antibiotics, call if not feeling better
- **HIGH Risk:** admit to hospital (≈10% of pneumonia patients die)

One goal was to compare various ML methods:
- logistic regression
- rule-based learning
- k-nearest neighbor
- neural nets
- Bayesian methods
- hierarchical mixtures of experts
- ...

Most accurate ML method: **multitask neural nets**

Safe to use neural nets on patients?

- **No — we used logistic regression instead...**
- **Why???”
Motivation: Predicting Pneumonia Risk Study (mid-90’s)

- **LOW Risk:** outpatient: antibiotics, call if not feeling better
- **HIGH Risk:** admit to hospital (≈10% of pneumonia patients die)

One goal was to compare various ML methods:
- logistic regression
- rule-based learning
- k-nearest neighbor
- neural nets
- Bayesian methods
- hierarchical mixtures of experts
- ...

Most accurate ML method: **multitask neural nets**

Safe to use neural nets on patients?

- No — we used logistic regression instead...
- Why???
Motivation: Predicting Pneumonia Risk Study (mid-90’s)

- **LOW Risk:** outpatient: antibiotics, call if not feeling better
- **HIGH Risk:** admit to hospital (≈10% of pneumonia patients die)

One goal was to compare various ML methods:
- logistic regression
- rule-based learning
- k-nearest neighbor
- neural nets
- Bayesian methods
- hierarchical mixtures of experts
- ...

Most accurate ML method: **multitask neural nets**

Safe to use neural nets on patients?

- **No — we used logistic regression instead…**
- Why???
Motivation: Predicting Pneumonia Risk Study (mid-90's)

- RBL learned rule: \textbf{HasAsthma}(x) \Rightarrow \textbf{LessRisk}(x)

- True pattern in data:
  - asthmatics presenting with pneumonia considered very high risk
  - receive aggressive treatment and often admitted to ICU
  - history of asthma also means they often go to healthcare sooner
  - treatment lowers risk of death compared to general population

- If RBL learned asthma is good for you, NN probably did, too
  - if we use NN for admission decision, could hurt asthmatics

- Key to discovering \textbf{HasAsthma}(x)... was intelligibility of rules
  - even if we can remove asthma problem from neural net, what other "bad patterns" don't we know about that RBL missed?
Motivation: Predicting Pneumonia Risk Study (mid-90’s)

- RBL learned rule: HasAsthma(x) $\Rightarrow$ LessRisk(x)

- True pattern in data:
  - asthmatics presenting with pneumonia considered very high risk
  - receive aggressive treatment and often admitted to ICU
  - history of asthma also means they often go to healthcare sooner
  - treatment lowers risk of death compared to general population

- If RBL learned asthma is good for you, NN probably did, too
  - if we use NN for admission decision, could hurt asthmatics

- Key to discovering HasAsthma(x)...
  - was intelligibility of rules
    - even if we can remove asthma problem from neural net, what other "bad patterns" don’t we know about that RBL missed?
Motivation: Predicting Pneumonia Risk Study (mid-90’s)

- RBL learned rule: HasAsthma(x) \implies\ LessRisk(x)

- True pattern in data:
  - asthmatics presenting with pneumonia considered very high risk
  - receive aggressive treatment and often admitted to ICU
  - history of asthma also means they often go to healthcare sooner
  - treatment lowers risk of death compared to general population

- If RBL learned asthma is good for you, NN probably did, too
  - if we use NN for admission decision, could hurt asthmatics

- Key to discovering HasAsthma(x)... was intelligibility of rules
  - even if we can remove asthma problem from neural net, what other ”bad patterns” don’t we know about that RBL missed?
RBL learned rule: \textbf{HasAsthma}(x) \implies \textbf{LessRisk}(x)

True pattern in data:
- asthmatics presenting with pneumonia considered very high risk
- receive aggressive treatment and often admitted to ICU
- history of asthma also means they often go to healthcare sooner
- treatment lowers risk of death compared to general population

If RBL learned asthma is good for you, NN probably did, too
- if we use NN for admission decision, could hurt asthmatics

Key to discovering \textbf{HasAsthma}(x)\ldots was intelligibility of rules
- even if we can remove asthma problem from neural net, what other "bad patterns" don’t we know about that RBL missed?
Lessons Learned

- Risky to use data for purposes it was not designed for
  - Most data has unexpected landmines
  - Not ethical to collect correct data for asthma

- Much too difficult to fully understand the data
  - Our approach is to make the learned models as intelligible as possible

- Must be able to understand models used in healthcare
  - Otherwise models can hurt patients because of true patterns in data
  - If you don’t understand model it will make bad mistakes

- Same story for race, gender, socioeconomic bias
  - The problem is in data and training signals, not learning algorithm
Lessons Learned

- Risky to use data for purposes it was not designed for
  - Most data has unexpected landmines
  - Not ethical to collect correct data for asthma

- Much too difficult to fully understand the data
  - Our approach is to make the learned models as intelligible as possible

- Must be able to understand models used in healthcare
  - Otherwise models can hurt patients because of true patterns in data
  - If you don’t understand model it will make bad mistakes

- Same story for race, gender, socioeconomic bias
  - The problem is in data and training signals, not learning algorithm
Lessons Learned

- Risky to use data for purposes it was not designed for
  - Most data has unexpected landmines
  - Not ethical to collect correct data for asthma

- Much too difficult to fully understand the data
  - Our approach is to make the learned models as intelligible as possible

- Must be able to understand models used in healthcare
  - Otherwise models can hurt patients because of true patterns in data
  - If you don’t understand model it will make bad mistakes

- Same story for race, gender, socioeconomic bias
  - The problem is in data and training signals, not learning algorithm
Lessons Learned

- Risky to use data for purposes it was not designed for
  - Most data has unexpected landmines
  - Not ethical to collect correct data for asthma

- Much too difficult to fully understand the data
  - Our approach is to make the learned models as intelligible as possible

- Must be able to understand models used in healthcare
  - Otherwise models can hurt patients because of true patterns in data
  - If you don’t understand model it will make bad mistakes

- Same story for race, gender, socioeconomic bias
  - The problem is in data and training signals, not learning algorithm
All we need is an accurate, intelligible model
Problem: The Accuracy vs. Intelligibility Tradeoff

Rich Caruana (Microsoft Research)
Problem: The Accuracy vs. Intelligibility Tradeoff

Intelligibility vs. Accuracy diagram:

- Boosted Trees
- Random Forests
- Neural Nets
- Single Decision Tree
- Logistic Regression
- Naive Bayes
- Decision Lists

Rich Caruana (Microsoft Research)
FAT/ML 2017: Intelligible Models
August 14, 2017
Model Space from Simple to Complex

- **Linear Model:** \( y = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n \)

- **Additive Model:** \( y = f_1(x_1) + ... + f_n(x_n) \)

- **Additive Model with Interactions:** \( y = \sum_i f_i(x_i) + \sum_{ij} f_{ij}(x_i, x_j) + \sum_{ijk} f_{ijk}(x_i, x_j, x_k) + ... \)

- **Full Complexity Model:** \( y = f(x_1, ..., x_n) \)
Model Space from Simple to Complex

- **Linear Model:** \[ y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n \]
- **Additive Model:** \[ y = f_1(x_1) + \ldots + f_n(x_n) \]
- **Additive Model with Interactions:** \[ y = \sum_i f_i(x_i) + \sum_{ij} f_{ij}(x_i, x_j) + \sum_{ijk} f_{ijk}(x_i, x_j, x_k) + \ldots \]
- **Full Complexity Model:** \[ y = f(x_1, \ldots, x_n) \]
Model Space from Simple to Complex

- **Linear Model:** \( y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n \)

- **Additive Model:** \( y = f_1(x_1) + \ldots + f_n(x_n) \)

- **Additive Model with Interactions:** \( y = \sum_i f_i(x_i) + \sum_{ij} f_{ij}(x_i, x_j) + \sum_{ijk} f_{ijk}(x_i, x_j, x_k) + \ldots \)

- **Full Complexity Model:** \( y = f(x_1, \ldots, x_n) \)
Model Space from Simple to Complex

- Linear Model: \( y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n \)
- Additive Model: \( y = f_1(x_1) + \ldots + f_n(x_n) \)
- Additive Model with Interactions: \( y = \sum_i f_i(x_i) + \sum_{ij} f_{ij}(x_i, x_j) + \sum_{ijk} f_{ijk}(x_i, x_j, x_k) + \ldots \)
- Full Complexity Model: \( y = f(x_1, \ldots, x_n) \)
Generalized Additive Models (GAMs)

- Developed at Stanford by Hastie and Tibshirani in late 80’s
- Regression: \( y = f_1(x_1) + \ldots + f_n(x_n) \)
- Classification: \( \logit(y) = f_1(x_1) + \ldots + f_n(x_n) \)
- Each feature is “shaped” by shape function \( f_i \)

T. Hastie and R. Tibshirani.  
*Generalized additive models.*  
Skip technical details of algorithm and jump to results
Motivation: Predicting Pneumonia Risk Study (mid-90’s)

- Pneumonia Data (dataset from early 1990’s)
  - 14,199 pneumonia patients
  - 70:30 train:test split (train=9847; test=4352)
  - 46 features
  - predict POD (probability of death)
  - 10.86% of patients (1542) died
### Physical examination findings

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Respiration rate (resps/min)</td>
<td>≤ 29*, ≥ 30</td>
</tr>
<tr>
<td>Heart rate (beats/min)</td>
<td>≤ 124*, 125–150, ≥ 151</td>
</tr>
<tr>
<td>Systolic blood pressure (mmHg)</td>
<td>≤ 60, 61–70, 71–80, 81–90, ≥ 91*</td>
</tr>
<tr>
<td>Temperature (°C)</td>
<td>≤ 34.4, 34.5–34.9, 35–35.5, 35.6–38.3*, 38.4–39.9, ≥ 40</td>
</tr>
<tr>
<td>Altered mental status (disorientation, lethargy, or coma)</td>
<td>no*, yes</td>
</tr>
<tr>
<td>Wheezing</td>
<td>no*, yes</td>
</tr>
<tr>
<td>Stridor</td>
<td>no*, yes</td>
</tr>
<tr>
<td>Heart murmur</td>
<td>no*, yes</td>
</tr>
<tr>
<td>Gastrointestinal bleeding</td>
<td>no*, yes</td>
</tr>
</tbody>
</table>

### Laboratory findings

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sodium level (mEq/l)</td>
<td>≤ 124, 125–130, 131–149*, ≥ 150</td>
</tr>
<tr>
<td>Potassium level (mEq/l)</td>
<td>≤ 5.2*, ≥ 5.3</td>
</tr>
<tr>
<td>Creatinine level (mg/dl)</td>
<td>≤ 1.6*, 1.7–3.0, 3.1–9.9, ≥ 10.0</td>
</tr>
<tr>
<td>Glucose level (mg/dl)</td>
<td>≤ 249*, 250–299, 300–399, ≥ 400</td>
</tr>
<tr>
<td>BUN level (mg/dl)</td>
<td>≤ 29*, 30 to 49, ≥ 50</td>
</tr>
<tr>
<td>Liver function tests (coded only as normal* or abnormal)</td>
<td>SGOT ≤ 63 and alkaline phosphatase ≤ 499*, SGOT &gt; 63 or alkaline phosphatase &gt; 499</td>
</tr>
<tr>
<td>Albumin level (gm/dl)</td>
<td>≤ 2.5, 2.6–3, ≥ 3.1*</td>
</tr>
<tr>
<td>Hematocrit</td>
<td>6–20, 20.1–24.9, 25–29, ≥ 30*</td>
</tr>
<tr>
<td>White blood cell count (1000 cells/μl)</td>
<td>0.1–3, 3.1–19.9*, ≥ 20</td>
</tr>
<tr>
<td>Percentage bands</td>
<td>≤ 10*, 11–20, 21–30, 31–50, ≥ 51</td>
</tr>
<tr>
<td>Blood pH</td>
<td>≤ 7.20, 7.21–7.35, 7.36–7.45*, ≥ 7.46</td>
</tr>
<tr>
<td>Blood pO₂ (mmHg)</td>
<td>≤ 59, 60–70, 71–75, ≥ 76*</td>
</tr>
<tr>
<td>Blood pCO₂ (mmHg)</td>
<td>≤ 44*, 45–55, 56–64, ≥ 65</td>
</tr>
</tbody>
</table>
What GA2Ms on Steroids Learn About Risk vs. Age

![Graph showing Pneumonia Risk Score and Density vs. Age]

- Pneumonia Risk Score
- Density
- Age

Rich Caruana (Microsoft Research)
Some of the things the intelligible model learned:

- Age 105 is safer than Age 95
- We should have a retirement variable
  - Has_Asthma => lower risk
  - History of chest pain => lower risk
  - History of heart disease => lower risk

- Good we didn't deploy neural net back in 1995
- But can understand, edit and safely deploy intelligible GA2M model
- Intelligible/transparent model is like having a magic pair of glasses

- Model correctness depends on how model will be used
  - this is a good model for health insurance providers
  - but needs to be repaired to use for hospital admissions

**Important:** Must keep potentially offending features in model!
Some of the things the intelligible model learned:

- Age 105 is safer than Age 95
- We should have a retirement variable
- Has_Asthma $\Rightarrow$ lower risk
- History of chest pain $\Rightarrow$ lower risk
- History of heart disease $\Rightarrow$ lower risk

Good we didn't deploy neural net back in 1995

But can understand, edit and safely deploy intelligible GA2M model

Intelligible/transparent model is like having a magic pair of glasses

Model correctness depends on how model will be used

- this is a good model for health insurance providers
- but needs to be repaired to use for hospital admissions

**Important:** Must keep potentially offending features in model!
Some of the things the intelligible model learned:

- Age 105 is safer than Age 95
- We should have a retirement variable
- Has_Asthma $\Rightarrow$ lower risk
- History of chest pain $\Rightarrow$ lower risk
- History of heart disease $\Rightarrow$ lower risk

Good we didn't deploy neural net back in 1995

But can understand, edit and safely deploy intelligible GA2M model

Intelligible/transparent model is like having a magic pair of glasses

Model correctness depends on how model will be used

- this is a good model for health insurance providers
- but needs to be repaired to use for hospital admissions

Important: Must keep potentially offending features in model!
Some of the things the intelligible model learned:
- Age 105 is safer than Age 95
- We should have a retirement variable
- Has_Asthma $\Rightarrow$ lower risk
- History of chest pain $\Rightarrow$ lower risk
- History of heart disease $\Rightarrow$ lower risk

Good we didn’t deploy neural net back in 1995
But can understand, edit and safely deploy intelligible GA2M model
Intelligible/transparent model is like having a magic pair of glasses

Model correctness depends on how model will be used
- this is a good model for health insurance providers
- but needs to be repaired to use for hospital admissions

Important: Must keep potentially offending features in model!
Some of the things the intelligible model learned:
- Age 105 is safer than Age 95
- We should have a retirement variable
- Has_Asthma $\Rightarrow$ lower risk
- History of chest pain $\Rightarrow$ lower risk
- History of heart disease $\Rightarrow$ lower risk

Good we didn’t deploy neural net back in 1995
But can understand, edit and safely deploy intelligible GA2M model
Intelligible/transparent model is like having a magic pair of glasses

Model correctness depends on how model will be used
- this is a good model for health insurance providers
- but needs to be repaired to use for hospital admissions

Important: Must keep potentially offending features in model!
Some of the things the intelligible model learned:

- Age 105 is safer than Age 95
- We should have a retirement variable
- Has_Asthma $\Rightarrow$ lower risk
- History of chest pain $\Rightarrow$ lower risk
- History of heart disease $\Rightarrow$ lower risk

Good we didn’t deploy neural net back in 1995

But can understand, edit and safely deploy intelligible GA2M model

Intelligible/transparent model is like having a magic pair of glasses

Model correctness depends on how model will be used

- this is a good model for health insurance providers
- but needs to be repaired to use for hospital admissions

Important: Must keep potentially offending features in model!
Parity is the classic (extreme) interaction
- For N-bit parity, need all N bits at the same time to calculate parity
- No correlation between any of the bits and parity signal
- No information in any subset of the bits

- Interactions can’t be modeled as sum of independent effects
- Interactions important on some problems, less on others
Age vs. Cancer Pairwise Interaction (Pneumonia-95)
30-day Hospital Readmission (joint work with NYP)

- 30-day Hospital Readmission Data
  - larger, modern dataset
  - records from NYP 2011-2014
  - train=195,901 (2011-12); test=100,823 (2013)
  - 3,956 features for each patient
  - goal: predict probability patient will be readmitted within 30 days
  - 8.91% of patients readmitted within 30 days
Age Plot for 30-Day Hospital Readmission
Age Plot for 30-Day Readmission: Zoom in on Age 0-2
Quick look at two 30-day Readmission Patients
Bias is in the data, not the learning algorithm

- Models trained on data will learn any biases in the data
  - ML for resume processing will learn gender bias if data has a gender bias
  - ML for recidivism prediction will learn race bias if data has a race bias
  - ...

How to deal with bias using transparent models:
- must keep bias features in data when model is trained
- remove what was learned from these bias features after training

If offending bias variables are eliminated prior to training:
- often can’t tell you still have a problem
- makes it harder to correct the problem

EU General Data Protection Regulation (goes into effect 2018):
- Article 9 makes it more difficult to use personal data revealing racial or ethnic origin and other “special categories”
Bias is in the data, not the learning algorithm

Models trained on data will learn any biases in the data
- ML for resume processing will learn gender bias if data has a gender bias
- ML for recidivism prediction will learn race bias if data has a race bias
- ...

How to deal with bias using transparent models:
- must keep bias features in data when model is trained
- remove what was learned from these bias features after training

If offending bias variables are eliminated prior to training:
- often can’t tell you still have a problem
- makes it harder to correct the problem

EU General Data Protection Regulation (goes into effect 2018):
- Article 9 makes it more difficult to use personal data revealing racial or ethnic origin and other “special categories”
Bias is in the data, not the learning algorithm

Models trained on data will learn any biases in the data
  - ML for resume processing will learn gender bias if data has a gender bias
  - ML for recidivism prediction will learn race bias if data has a race bias
  - ...

How to deal with bias using transparent models:
  - must keep bias features in data when model is trained
  - remove what was learned from these bias features after training

If offending bias variables are eliminated prior to training:
  - often can’t tell you still have a problem
  - makes it harder to correct the problem

EU General Data Protection Regulation (goes into effect 2018):
  - Article 9 makes it more difficult to use personal data revealing racial or ethnic origin and other “special categories”
Bias is in the data, not the learning algorithm
Models trained on data will learn any biases in the data
  - ML for resume processing will learn gender bias if data has a gender bias
  - ML for recidivism prediction will learn race bias if data has a race bias
  - ...

How to deal with bias using transparent models:
  - must keep bias features in data when model is trained
  - remove what was learned from these bias features after training

If offending bias variables are eliminated prior to training:
  - often can’t tell you still have a problem
  - makes it harder to correct the problem

EU General Data Protection Regulation (goes into effect 2018):
  - Article 9 makes it more difficult to use personal data revealing racial or ethnic origin and other “special categories”
Bias is in the data, not the learning algorithm

- Models trained on data will learn any biases in the data
  - ML for resume processing will learn gender bias if data has a gender bias
  - ML for recidivism prediction will learn race bias if data has a race bias
  - ...

How to deal with bias using transparent models:
- must keep bias features in data when model is trained
- remove what was learned from these bias features after training

If offending bias variables are eliminated prior to training:
- often can’t tell you still have a problem
- makes it harder to correct the problem

EU General Data Protection Regulation (goes into effect 2018):
- Article 9 makes it more difficult to use personal data revealing racial or ethnic origin and other “special categories”
But we have yet another transparency trick up our sleeve...
Is COMPAS model biased?
Is recidivism training data biased?
Is COMPAS model more biased than training data might warrant?

Transparent Modeling Trick:
- train transparent model #1 on raw recidivism data
- train transparent model #2 on COMPAS model predictions
- Real vs. Memorex:
  compare what is learned in model #1 from raw data to model #2 from COMPAS mimic
- one caveat: ProPublica data doesn't have all of the COMPAS features
ProPublica COMPAS Recidivism Data

- Is COMPAS model biased?
- Is recidivism training data biased?
- Is COMPAS model more biased than training data might warrant?

Transparent Modeling Trick:
- train transparent model #1 on raw recidivism data
- train transparent model #2 on COMPAS model predictions
- Real vs. Memorex:
  compare what is learned in model #1 from raw data to model #2 from COMPAS mimic

- one caveat: ProPublica data doesn't have all of the COMPAS features
Is COMPAS model biased?
Is recidivism training data biased?
Is COMPAS model more biased than training data might warrant?

Transparent Modeling Trick:
- train transparent model #1 on raw recidivism data
- train transparent model #2 on COMPAS model predictions
- Real vs. Memorex:
  compare what is learned in model #1 from raw data to model #2 from COMPAS mimic
- one caveat: ProPublica data doesn't have all of the COMPAS features
Advantages of Transparent Modeling for Bias Detection and Elimination

- Don’t have to know what biases to look for in advance
- Don’t have to design statistical tests for biases in advance
- Just train model, and look at what it learned — model can (and will!) surprise you

- Modularity of GAMs makes bias easier to recognize
- Modularity of GAMs makes bias easier to correct

- Accuracy of GA2Ms means less is missing — GA2M model often is as accurate as any other model black-box we could train on data
Advantages of Transparent Modeling for Bias Detection and Elimination

- Don’t have to know what biases to look for in advance
- Don’t have to design statistical tests for biases in advance
- Just train model, and look at what it learned — model can (and will!) surprise you

- Modularity of GAMs makes bias easier to recognize
- Modularity of GAMs makes bias easier to correct

- Accuracy of GA2Ms means less is missing — GA2M model often is as accurate as any other model black-box we could train on data
Advantages of Transparent Modeling for Bias Detection and Elimination

- Don’t have to know what biases to look for in advance
- Don’t have to design statistical tests for biases in advance
- Just train model, and look at what it learned — model can (and will!) surprise you

- Modularity of GAMs makes bias easier to recognize
- Modularity of GAMs makes bias easier to correct

- Accuracy of GA2Ms means less is missing — GA2M model often is as accurate as any other model black-box we could train on data
• **GA2Ms are not causal models**
  - because they’re simple and transparent, often find causal effects
  - but it’s up to the user to figure out what’s really going on

• **GA2Ms do not cure the curse of dimensionality and correlation**

• **GA2Ms are intelligible only if features are intelligible**

• **GA2Ms are not a replacement for deep learning on raw signals**
  - does not work as well as deep nets on pixels, speech signals, ...
  - works best on features crafted by humans

• **GA2Ms are not perfect yet...**
  - we’re still doing research to make the GA2Ms better
  - but they’re now good enough to be used instead of logistic regression
GA2Ms are not causal models
- because they’re simple and transparent, often find causal effects
- but it’s up to the user to figure out what’s really going on

GA2Ms do not cure the curse of dimensionality and correlation
- GA2Ms are intelligible only if features are intelligible
- GA2Ms are not a replacement for deep learning on raw signals
  - does not work as well as deep nets on pixels, speech signals, ...
  - works best on features crafted by humans

GA2Ms are not perfect yet...
- we’re still doing research to make the GA2Ms better
- but they’re now good enough to be used instead of logistic regression
GA2Ms are not causal models
because they’re simple and transparent, often find causal effects
but it’s up to the user to figure out what’s really going on

GA2Ms do not cure the curse of dimensionality and correlation

GA2Ms are intelligible only if features are intelligible
GA2Ms are not a replacement for deep learning on raw signals
  does not work as well as deep nets on pixels, speech signals, ...
  works best on features crafted by humans

GA2Ms are not perfect yet...
we’re still doing research to make the GA2Ms better
but they’re now good enough to be used instead of logistic regression
GA2Ms are not causal models
- because they’re simple and transparent, often find causal effects
- but it’s up to the user to figure out what’s really going on

GA2Ms do not cure the curse of dimensionality and correlation

GA2Ms are intelligible only if features are intelligible
- does not work as well as deep nets on pixels, speech signals, ...
- works best on features crafted by humans

GA2Ms are not perfect yet...
- we’re still doing research to make the GA2Ms better
- but they’re now good enough to be used instead of logistic regression
High accuracy on test set is not enough
There are land mines hidden in the data
You need magic glasses to see the landmines
It’s critical to understand model before deploying it
Model correctness depends on how model will be used
New GA2Ms give us accuracy and intelligibility at same time
Important to keep potentially offending variables in model so bias can be detected and then removed after training
If you eliminate offending variables before training you:
  - can’t tell you have a problem
  - make it harder to correct the problem
Transparency allows you to detect problems you didn’t anticipate in advance
References

Y. Lou, R. Caruana, and J. Gehrke.  
*Intelligible Models for Classification and Regression.*  
In *KDD*, 2012.

*Accurate Intelligible Models With Pairwise Interactions.*  
In *KDD*, 2013.

*Intelligible Models for Healthcare.*  
In *KDD*, 2015.

T. Hastie and R. Tibshirani.  
*Generalized additive models.*  
Thank You!

Thank you to MSR, after 5+ years of research we now have something no one else has that’s important for healthcare and also for regulatory transparency and dealing with bias. The world is beginning to take note and this work already has been mentioned in half a dozen press articles.
Stage 1: build best additive model using only 1-dim components
   - Additive effects are now modeled
   - If Stage 1 done perfectly, only have interaction (and noise) in residuals

Stage 2: fix the one-dimensional functions
   - Detect pairwise interactions on residuals (new FAST algorithm)

Stage 3: build shape models for most important pairwise interactions on residuals

Stage 4: post-process shape plots
   - center average prediction of each plot to improve modularity
   - sort terms by importance to aid intelligibility

Bag (repeat) process 10-100 times to create pseudo-confidence intervals and further reduce overfitting