Efficient Discovery of Heterogeneous Treatment Effects via Anomalous Pattern Detection

Edward McFowland III

emcfowla@umn.edu

Joint work with (Sriram Somanchi & Daniel B. Neill)

University of Minnesota

Driven to Discover[SM]

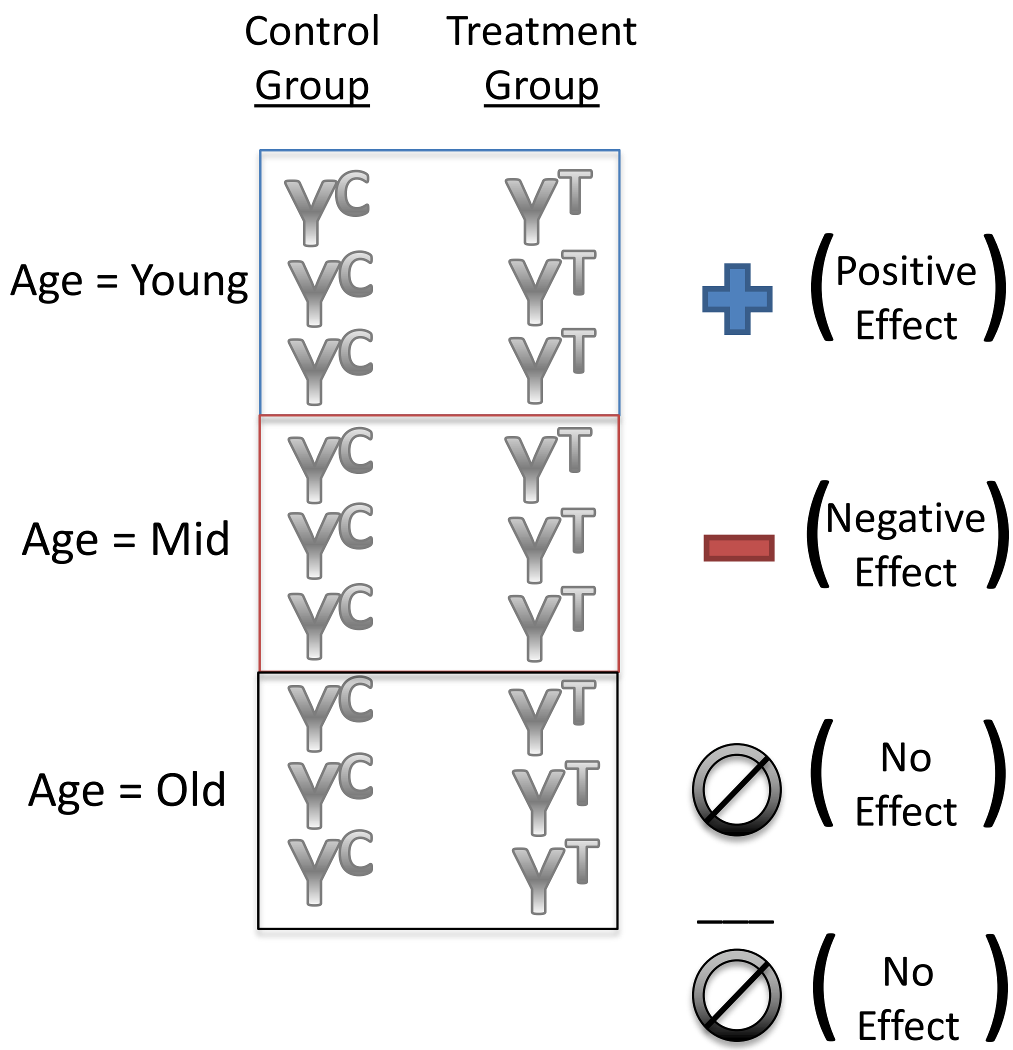
Agenda

- Introduction and Motivation
- Machine Learning’s Contributions (and Limits)
- How Anomaly Detection Can Help
- Treatment Effect Subset Scan
  - Algorithm
  - Statistical Properties
- Results
- Conclusions

Treatment Effect Heterogeneity

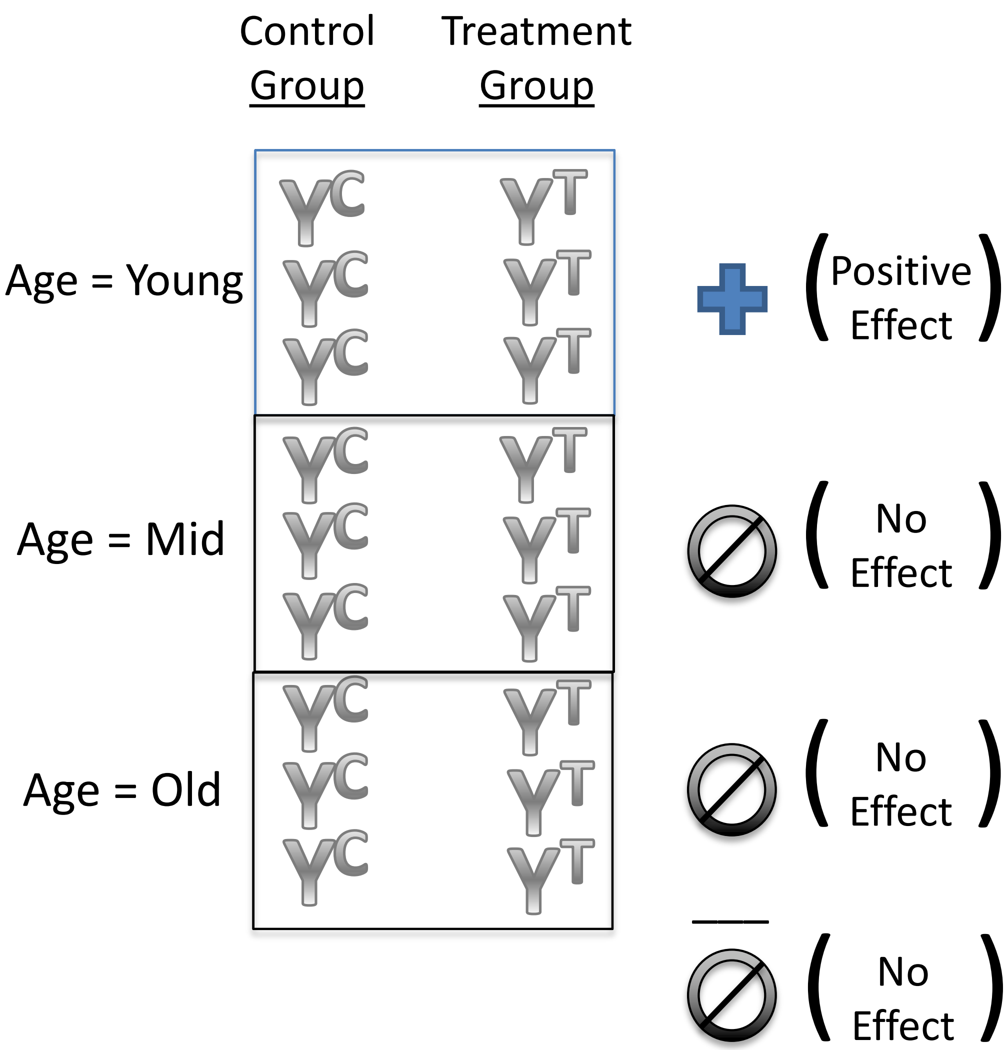
	Control <u>Group</u>	Treatment <u>Group</u>
Age = Young	Y^C Y^C Y^C	Y^T Y^T Y^T
Age = Mid	Y^C Y^C Y^C	Y^T Y^T Y^T
Age = Old	Y^C Y^C Y^C	Y^T Y^T Y^T

Treatment Effect Heterogeneity



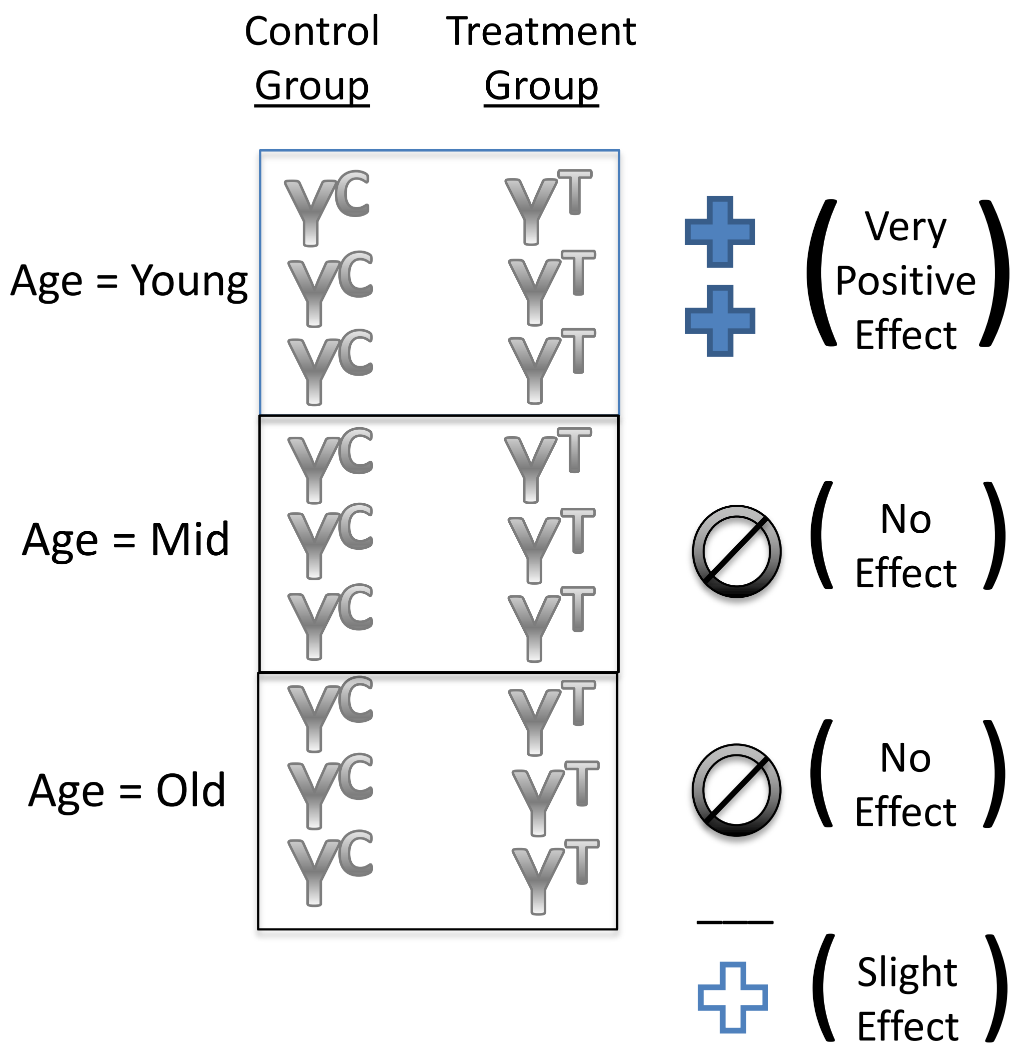
- Positive and negative effects can cancel

Treatment Effect Heterogeneity



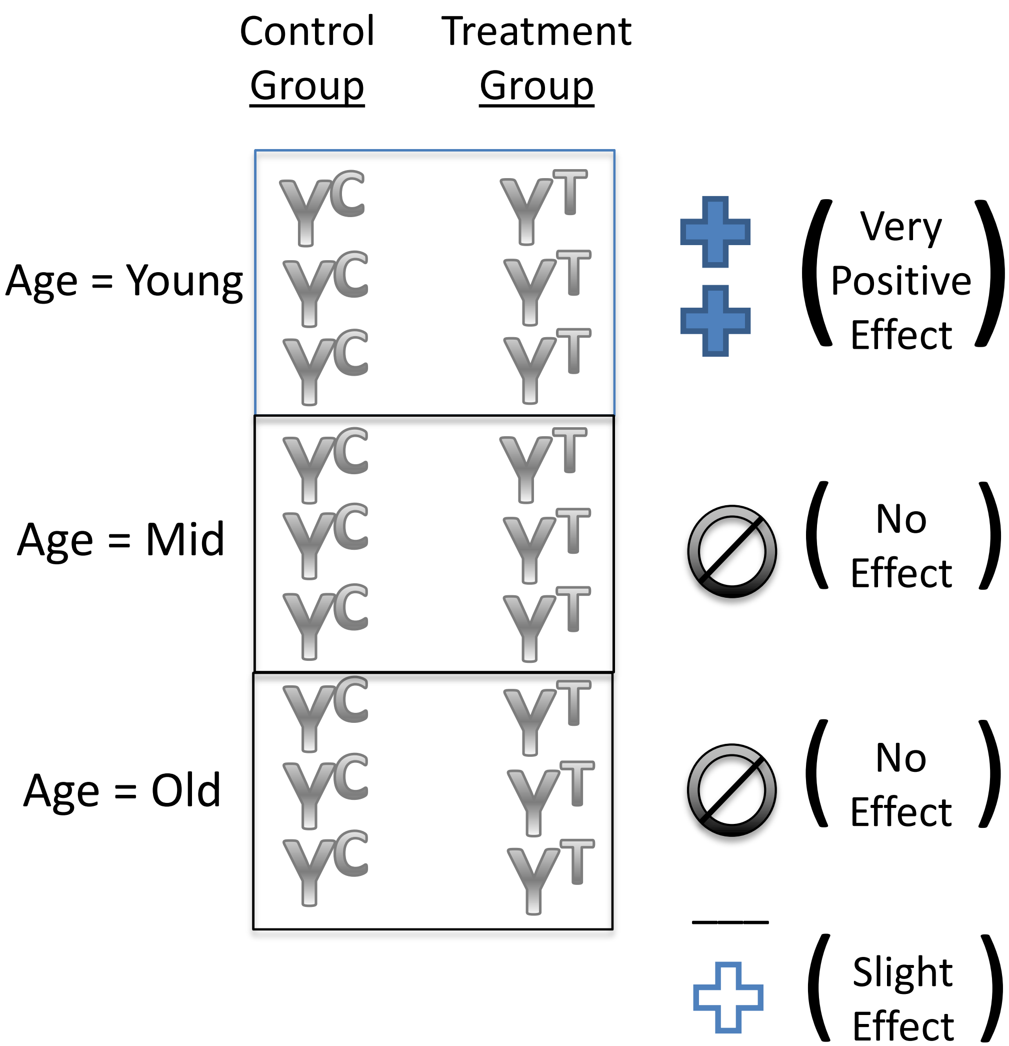
- Positive and negative effects can cancel
- True effect can be masked

Treatment Effect Heterogeneity



- Positive and negative effects can cancel
- True effect can be masked
- Effects could really be driven by a subpopulation

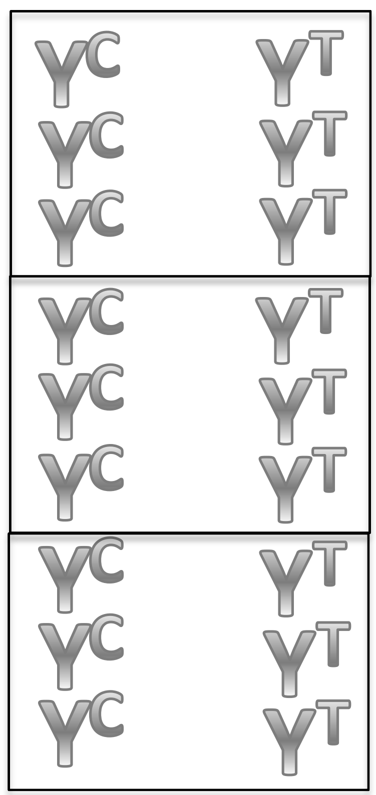
Treatment Effect Heterogeneity



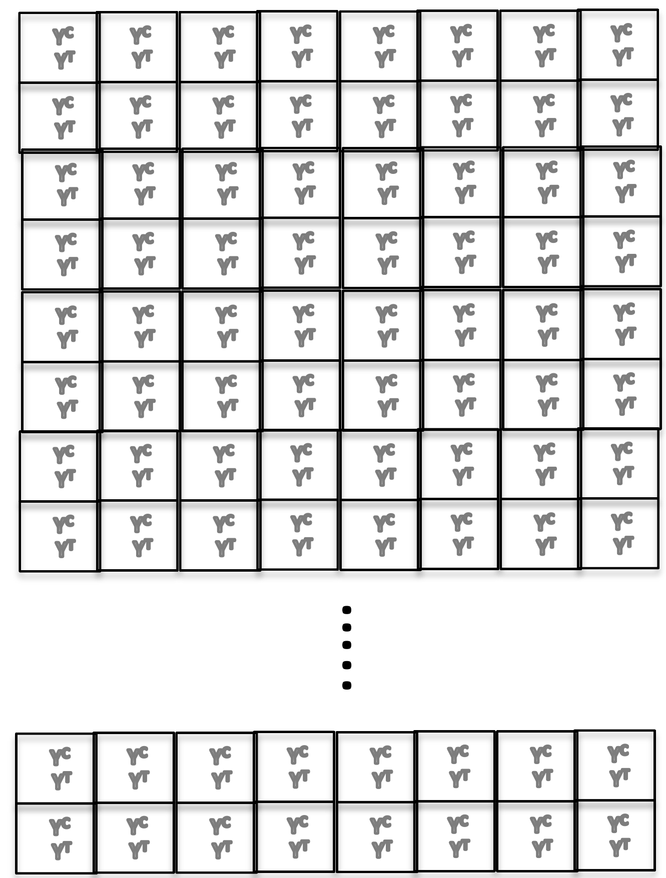
- Positive and negative effects can cancel
- True effect can be masked
  - Ex: FDA Approved BiDil Drug
- Effects could really be driven by a subpopulation
  - Ex: Perry Preschool

Treatment Effect Heterogeneity

Control Group | Treatment Group



Bing → Google



Agenda

- Introduction and Motivation
- Machine Learning's Contributions (and Limits)
- How Anomaly Detection Can Help
- Treatment Effect Subset Scan
  - Algorithm
  - Statistical Properties
- Results
- Conclusions

Machine Learning’s Contributions

- Regression Methods
  - OLS and Regularized Regression (e.g., LASSO)*
  - Imai and Ratkovic (2013)
  - Taddy et al. (2016)

- Single Tree Methods
  - Su et al (2009)
  - Imai and Strauss (2011)
  - Athey and Imbens (2017)*

- Ensemble Methods
  - Grimmer et al. (2017)
  - Wager and Athey (2017)*

* provide frequentist asym confidence intervals, for inference of effects significance

Limitations

- Regression Methods
  - Pre-specification of the model
- Single Tree Methods
  - Greedy and unstable
- Ensemble Methods
  - Fairly uninterpretable/no natural subpopulations

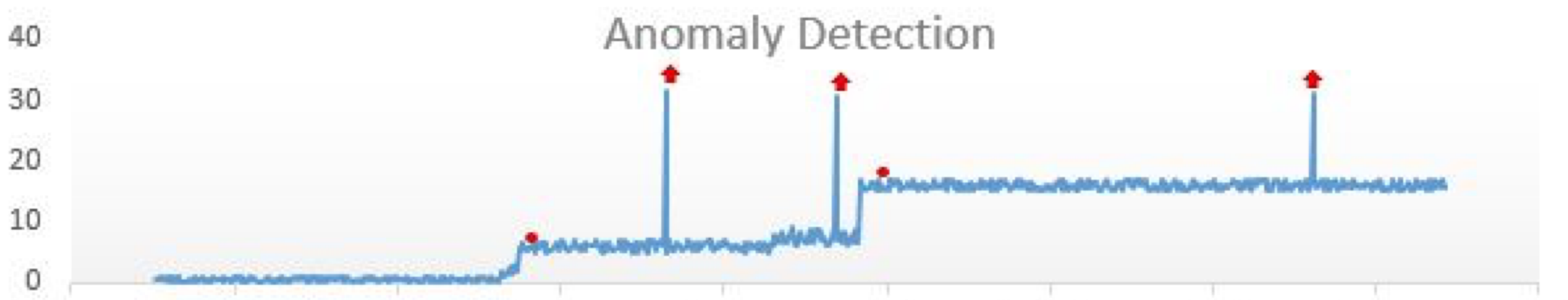
- General Limitations
  - The mean and only the mean
    - Other moments can be effected
    - Simpsons Paradox
  - Risk minimization not effect maximization
    - Small number of subpopulations considered
    - No guarantee on their "interestingness"
  - No "discovery", only model inspection

Agenda

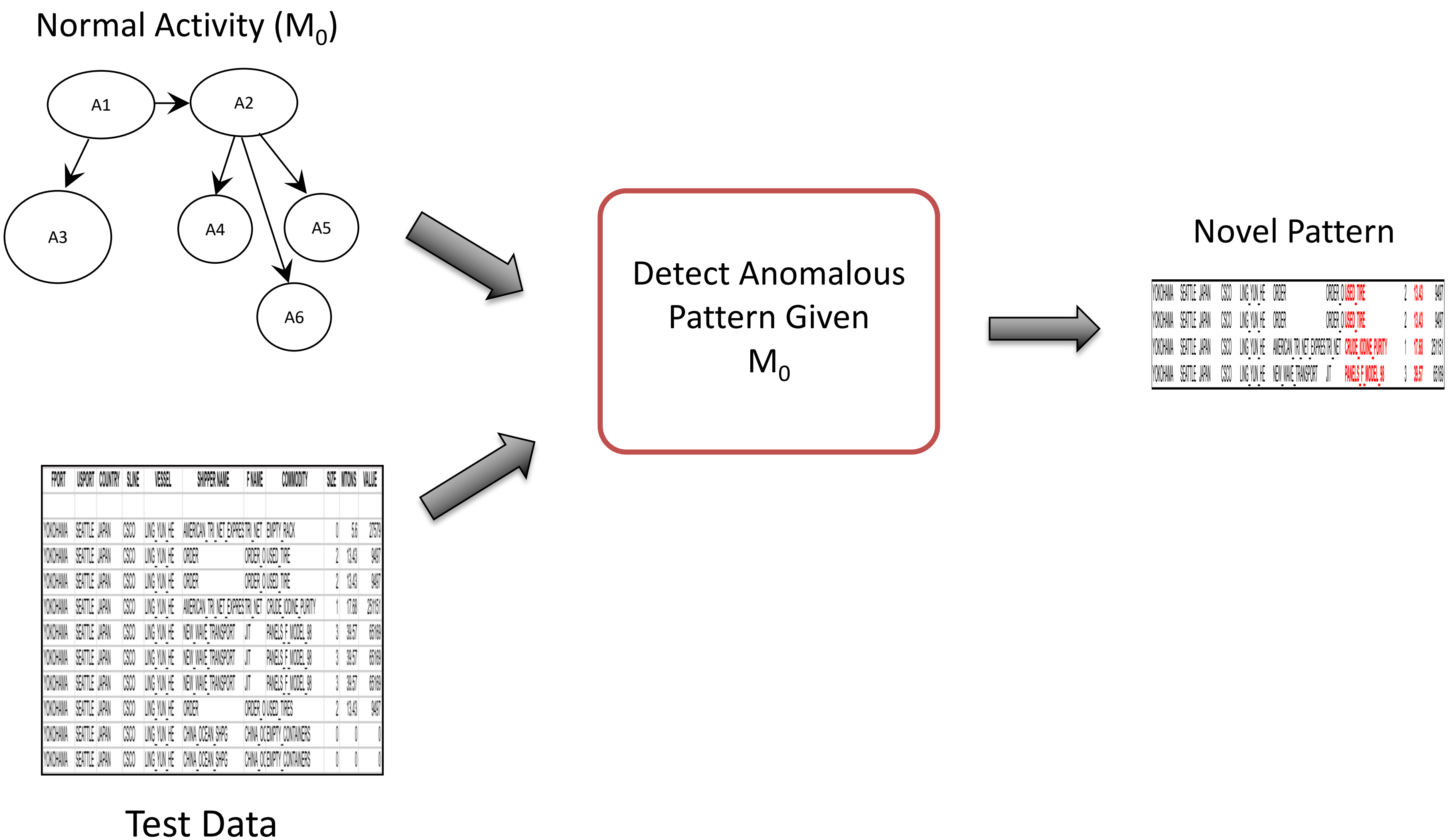
- Introduction and Motivation
- Machine Learning’s Contributions (and Limits)
- How Anomaly Detection Can Help
- Treatment Effect Subset Scan
  - Algorithm
  - Statistical Properties
- Results
- Conclusions

Anomaly Detection Paradigm

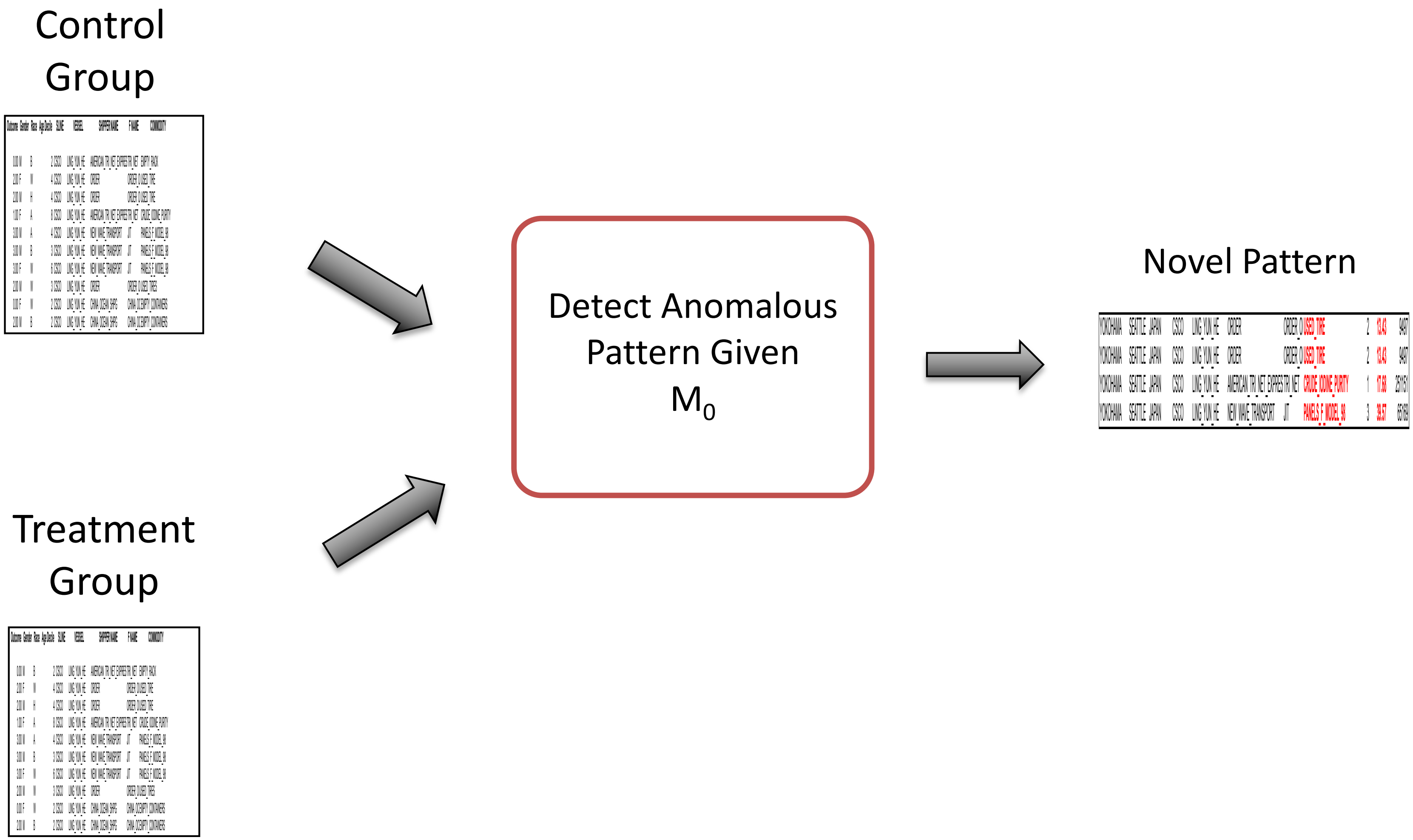
- Identifying when a "system" deviates away from its expected behavior.



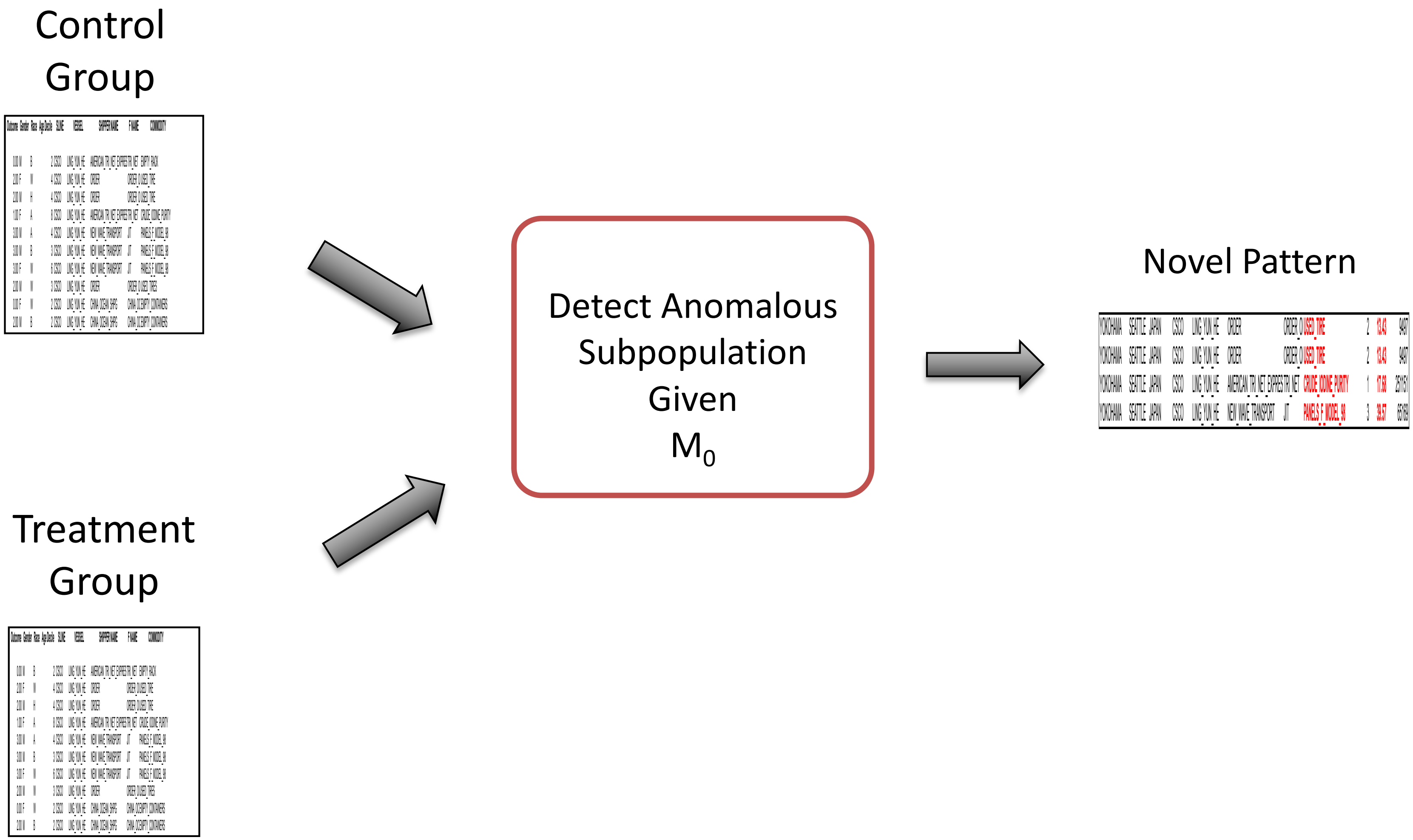
Anomalous Pattern Detection Procedure



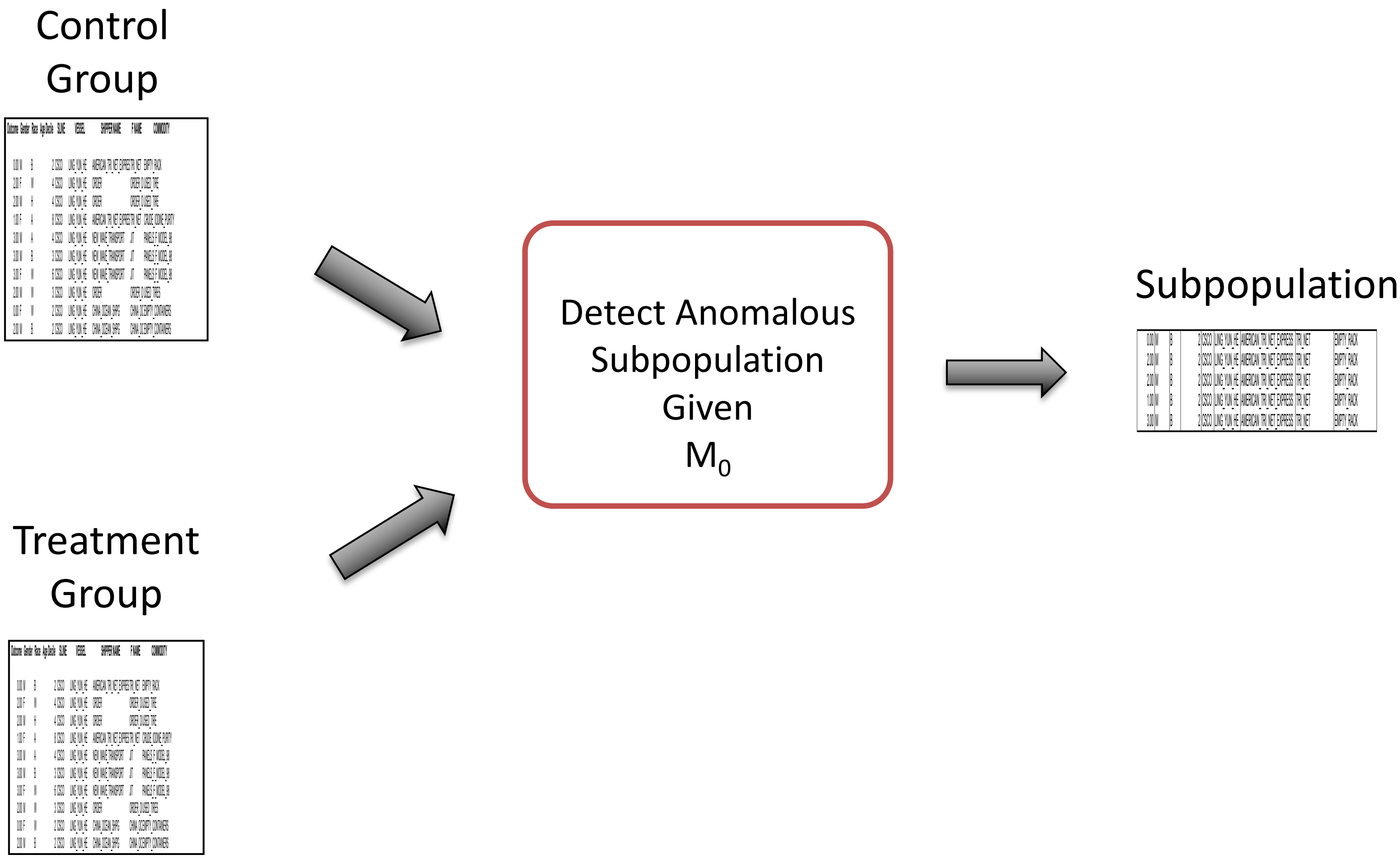
HTE Pattern Detection



HTE Pattern Detection



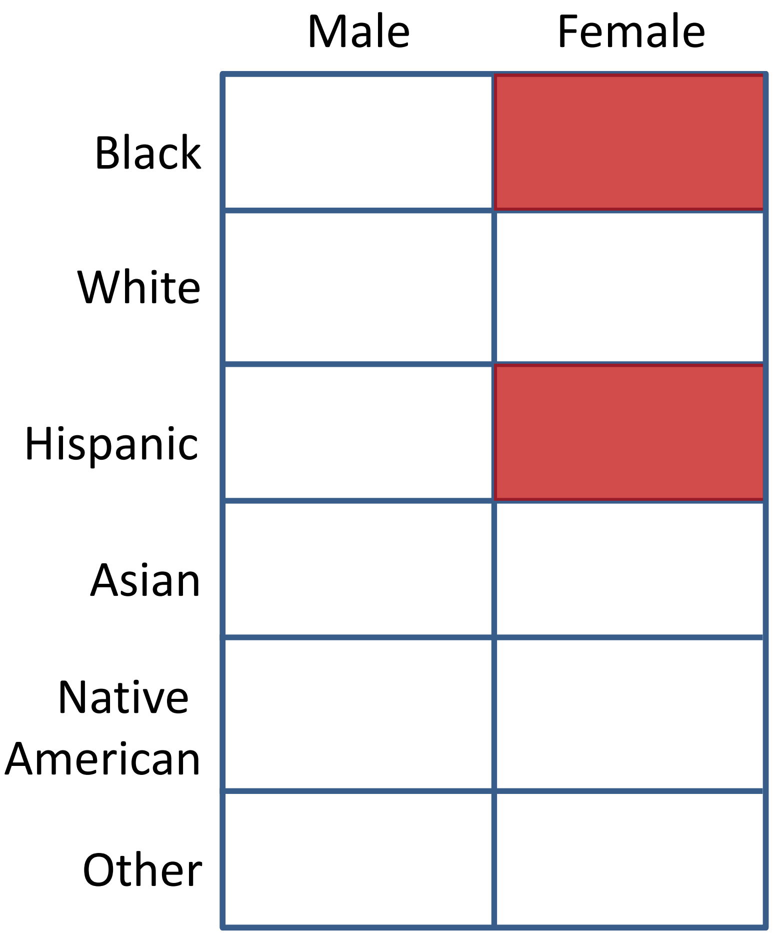
HTE Pattern Detection



Agenda

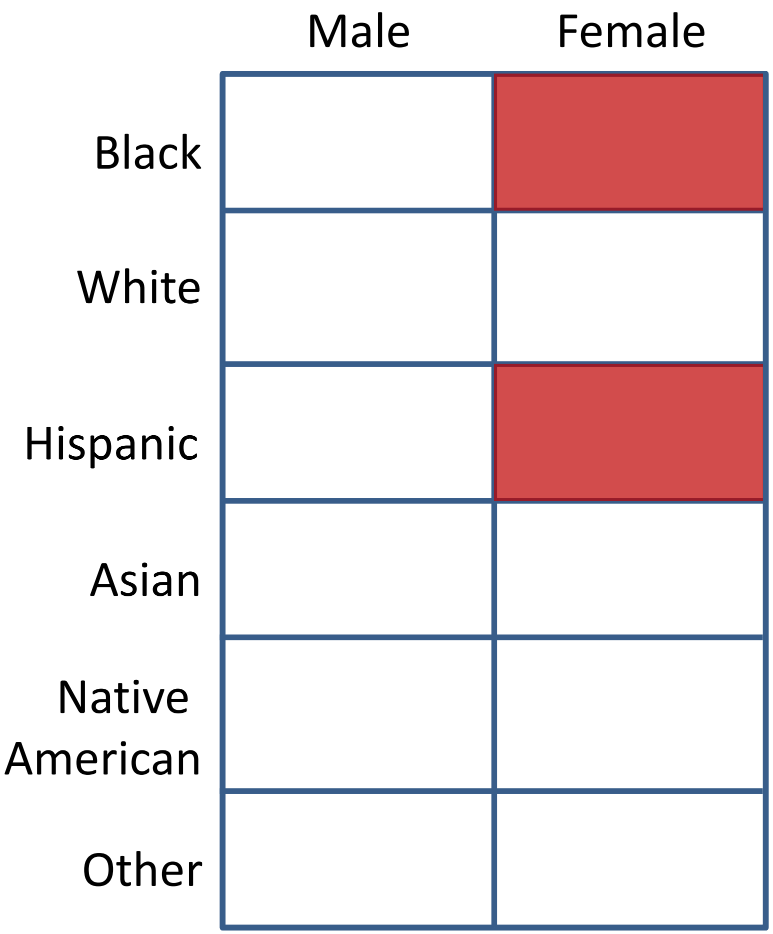
- Introduction and Motivation
- Machine Learning’s Contributions (and Limits)
- How Anomaly Detection Can Help
- Treatment Effect Subset Scan
  - Algorithm
  - Statistical Properties
- Results
- Conclusions

The Goal



Detect a subpopulation (subsets of attribute values), which correspond to anomalous outcomes for subjects in the treatment group

The Goal



Detect a subpopulation (subsets of attribute values), which correspond to anomalous outcomes for subjects in the treatment group

The Optimization

$s_1 \subseteq \{a_1 \ldots a_t\}, \ldots , s_M \subseteq \{a_1 \ldots a_p\}$

The Goal

	Male	Female
Black		
White		
Hispanic		
Asian		
Native American		
Other		

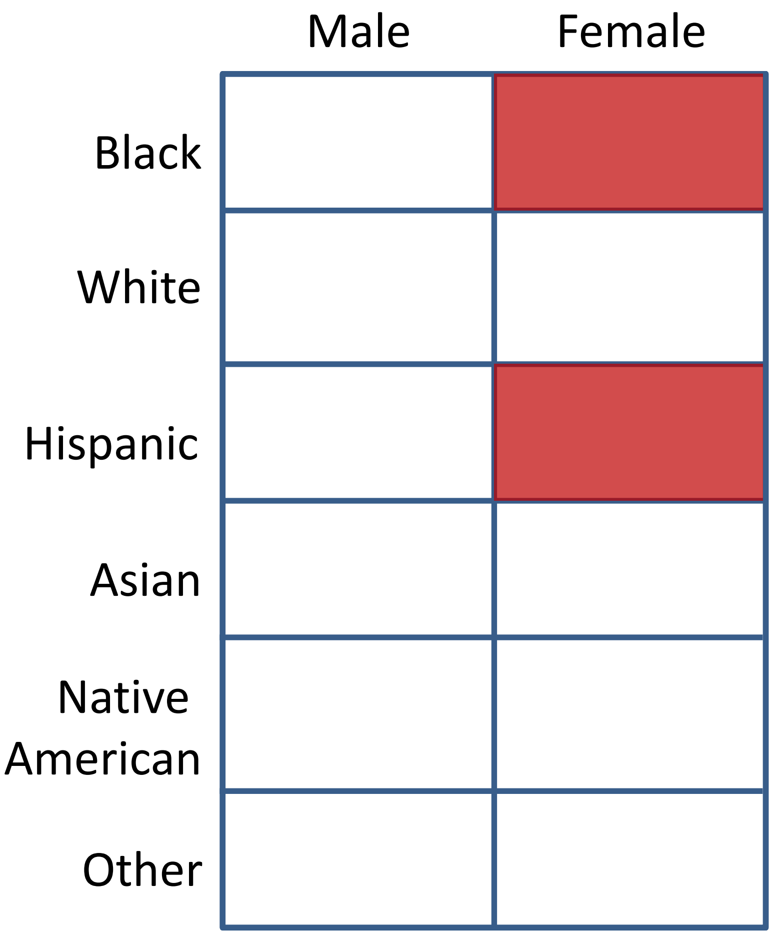
Detect a subpopulation (subsets of attribute values), which correspond to anomalous outcomes for subjects in the treatment group

<u>The Optimization</u>

$$s_1 \subseteq \{a_1 \ldots a_t\}, \ldots, s_M \subseteq \{a_1 \ldots a_p\}$$

$$S = s_1 \times \ldots \times s_M$$

The Goal



Detect a subpopulation (subsets of attribute values), which correspond to anomalous outcomes for subjects in the treatment group

<u>The Optimization</u>

$$s_1 \subseteq \{a_1 \ldots a_t\}, \ldots, s_M \subseteq \{a_1 \ldots a_p\}$$

$$S = s_1 \times \ldots \times s_M$$

$$S^* = \mathit{argmax}_S\, F(S)$$

Treatment Effects Subset Scan (TESS)

	Male	Female
Black	Y^{BM}	Y^{BF}
White	Y^{WM}	Y^{WF}
Hispanic	Y^{HM}	Y^{HF}
Asian	Y^{AM}	Y^{AF}
Native American	Y^{NM}	Y^{NF}
Other	Y^{OM}	Y^{OF}

I. Compute the statistical anomalousness of each treatment group subject

II. Detect subpopulation that is collectively the most anomalous

Treatment Effects Subset Scan (TESS)

	Male	Female
Black	P^{BM}	P^{BF}
White	P^{WM}	P^{WF}
Hispanic	P^{HM}	P^{HF}
Asian	P^{AM}	P^{AF}
Native American	P^{NM}	P^{NF}
Other	P^{OM}	P^{OF}

I. Compute the statistical anomalousness of each treatment group subject

-- This measurement will be a p-value

II. Detect subpopulation that is collectively the most anomalous

-- Many subjects with significant p-values

Treatment Effects Subset Scan (TESS)

	Male	Female
Black	Y^{BM}	Y^{BF}
White	Y^{WM}	Y^{WF}
Hispanic	Y^{HM}	Y^{HF}
Asian	Y^{AM}	Y^{AF}
Native American	Y^{NM}	Y^{NF}
Other	Y^{OM}	Y^{OF}

Control Group

I. Compute the statistical anomalousness of each treatment group subject

1. Estimate Conditional Distribution Under H_0

Treatment Effects Subset Scan (TESS)

	Male	Female
Black	Y^{BM}	Y^{BF}
White	Y^{WM}	Y^{WF}
Hispanic	Y^{HM}	Y^{HF}
Asian	Y^{AM}	Y^{AF}
Native American	Y^{NM}	Y^{NF}
Other	Y^{OM}	Y^{OF}

Treatment Group

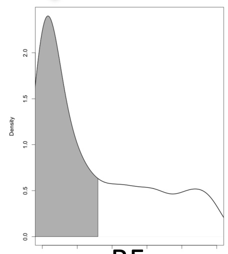
I. Compute the statistical anomalousness of each treatment group subject

1. Estimate Conditional Distribution Under H_0
2. Compute empirical p-values

Treatment Effects Subset Scan (TESS)

	Male	Female
Black	Y^{BM}	Y^{BF}
White	Y^{WM}	Y^{WF}
Hispanic	Y^{HM}	Y^{HF}
Asian	Y^{AM}	Y^{AF}
Native American	Y^{NM}	Y^{NF}
Other	Y^{OM}	Y^{OF}

Treatment Group



y_i^{BF}

I. Compute the statistical anomalousness of each treatment group subject

1. Estimate Conditional Distribution Under H_0
2. Compute empirical p-values

Treatment Effects Subset Scan (TESS)

	Male	Female
Black	P^{BM}	P^{BF}
White	P^{WM}	P^{WF}
Hispanic	P^{HM}	P^{HF}
Asian	P^{AM}	P^{AF}
Native American	P^{NM}	P^{NF}
Other	P^{OM}	P^{OF}

Treatment Group

I. Compute the statistical anomalousness of each treatment group subject

1. Estimate Conditional Distribution Under H_0
2. Compute empirical p-values

Treatment Effects Subset Scan (TESS)

	Male	Female
Black	P^{BM}	P^{BF}
White	P^{WM}	P^{WF}
Hispanic	P^{HM}	P^{HF}
Asian	P^{AM}	P^{AF}
Native American	P^{NM}	P^{NF}
Other	P^{OM}	P^{OF}

Treatment Group

I. Compute the statistical anomalousness of each treatment group subject

1. Estimate Conditional Distribution Under H_0
2. Compute empirical p-values
   i. Maps each bin's distribution to the same interval

Treatment Effects Subset Scan (TESS)

	Male	Female
Black	P^{BM}	P^{BF}
White	P^{WM}	P^{WF}
Hispanic	P^{HM}	P^{HF}
Asian	P^{AM}	P^{AF}
Native American	P^{NM}	P^{NF}
Other	P^{OM}	P^{OF}

Treatment Group

I. Compute the statistical anomalousness of each treatment group subject

1. Estimate Conditional Distribution Under H_0
2. Compute empirical p-values
   - i. Maps each bin's distribution to the same interval
   - ii. $P^{ij} \sim \text{Uniform}[0,1]$ under H_0

Treatment Effects Subset Scan (TESS)

	Male	Female
Black	P^{BM}	P^{BF}
White	P^{WM}	P^{WF}
Hispanic	P^{HM}	P^{HF}
Asian	P^{AM}	P^{AF}
Native American	P^{NM}	P^{NF}
Other	P^{OM}	P^{OF}

Treatment Group

I. Compute the statistical anomalousness of each treatment group subject

1. Estimate Conditional Distribution Under H_0
2. Compute empirical p-values
   i. Maps each bin's distribution to the same interval
   ii. P^{ij} ~ Uniform[0,1] under H_0
   iii. For any N p-values, we expect $N*\alpha$ to be significant at level α

Treatment Effects Subset Scan (TESS)

	Male	Female
Black	P^{BM}	P^{BF}
White	P^{WM}	P^{WF}
Hispanic	P^{HM}	P^{HF}
Asian	P^{AM}	P^{AF}
Native American	P^{NM}	P^{NF}
Other	P^{OM}	P^{OF}

Treatment Group

I. Compute the statistical anomalousness of each treatment group subject

1. Estimate Conditional Distribution Under H_0
2. Compute empirical p-values
   - i. Maps each bin's distribution to the same interval
   - ii. $P^{ij} \sim$ Uniform[0,1] under H_0
   - iii. For any N p-values, we expect $N*\alpha$ to be significant at level α

Higher Criticism:

$$F(S) = \max_{\alpha} \frac{N_{\alpha} - N\alpha}{\sqrt{N\alpha(1-\alpha)}}$$

Treatment Effects Subset Scan (TESS)

	Male	Female
Black	P^{BM}	P^{BF}
White	P^{WM}	P^{WF}
Hispanic	P^{HM}	P^{HF}
Asian	P^{AM}	P^{AF}
Native American	P^{NM}	P^{NF}
Other	P^{OM}	P^{OF}

Treatment Group

I. Compute the statistical anomalousness of each treatment group subject

II. Discover subsets of attribute values that define the most anomalous outcomes

1. Maximize F(S) over all subsets of $s_1 \times \ldots \times s_M$
   - Naïve search is infeasible $O(2^{\Sigma|A_i|})$

Treatment Effects Subset Scan (TESS)

Nonparametric Scan Statistic (NPSS)

Have: $S \subseteq \{A_1 \times \ldots \times A_M\}$
$= \{s_1 \times \ldots \times s_M\}$

I. Compute the statistical anomalousness of each treatment group subject

II. Discover subsets of attribute values that define the most anomalous outcomes

1. Maximize F(S) over all subsets of $s_1 \times \ldots \times s_M$
   - Naïve search is infeasible $O(2^{\Sigma |A_i|})$

Treatment Effects Subset Scan (TESS)

Nonparametric Scan Statistic (NPSS)

Have: $S \subseteq \{A_1 \times \ldots \times A_M\}$
$= \{s_1 \times \ldots \times s_M\}$

Select: $F(S) = max_\alpha \, \phi(\alpha, N_\alpha(S), N(S))$

Want: $max_S \, F(S)$

Assume: $\phi \uparrow$ w.r.t N_α
$\phi \downarrow$ w.r.t N and α
ϕ is convex

I. Compute the statistical anomalousness of each treatment group subject

II. Discover subsets of attribute values that define the most anomalous outcomes

1. Maximize F(S) over all subsets of $s_1 \times \ldots \times s_M$
   - Naïve search is infeasible $O(2^{\Sigma|A_i|})$

Treatment Effects Subset Scan (TESS)

Nonparametric Scan Statistic (NPSS)

Have: $S \subseteq \{A_1 \times \ldots \times A_M\}$
$= \{s_1 \times \ldots \times s_M\}$

Select: $F(S) = max_\alpha\, \phi(\alpha, N_\alpha(S), N(S))$

Want: $max_S\, F(S)$

Assume: $\phi \uparrow$ w.r.t N_α
$\phi \downarrow$ w.r.t N and α
ϕ is convex

There Exist: $G(a_i)$

Such That: $\max_{s_j \subseteq \{a_1, \ldots, a_t\}} F(s_j | A_{-j}) = \max_{i=1\ldots t} F\left(\{a_{(1)} \ldots a_{(t)}\} | A_{-j}\right)$

Only Consider: $\{a_{(1)}\}$
$\{a_{(1)}, a_{(2)}\}$
$\vdots$
$\{a_{(1)}, \ldots\ldots\ldots\ldots, a_{(M)}\}$

I. Compute the statistical anomalousness of each treatment group subject

II. Discover subsets of attribute values that define the most anomalous outcomes

1. Maximize F(S) over all subsets of $s_1 \times \ldots \times s_M$
   - Naïve search is infeasible $O(2^{\Sigma|A_i|})$

Treatment Effects Subset Scan (TESS)

Nonparametric Scan Statistic (NPSS)

Have: $S \subseteq \{A_1 \times \ldots \times A_M\}$
$= \{s_1 \times \ldots \times s_M\}$

Select: $F(S) = max_\alpha\ \phi(\alpha, N_\alpha(S), N(S))$

Want: $max_S\ F(S)$

Assume: $\phi \uparrow$ w.r.t N_α
$\phi \downarrow$ w.r.t N and α
ϕ is convex

There Exist: $G(a_i)$

Such That: $\max_{s_j \subseteq \{a_1, \ldots, a_t\}} F(s_j | A_{-j}) = \max_{i=1\ldots t} F\left(\left\{a_{(1)} \ldots a_{(t)}\right\} | A_{-j}\right)$

Only Consider: $\{$Black$\}$
$\{$Black, Hispanic$\}$
⋮
$\{$Black, Hispanic, Asian, …, White $\}$

I. Compute the statistical anomalousness of each treatment group subject

II. Discover subsets of attribute values that define the most anomalous outcomes

1. Maximize F(S) over all subsets of $s_1 \times \ldots \times s_M$
   - Naïve search is infeasible $O(2^{\Sigma|A_i|})$

Treatment Effects Subset Scan (TESS)

Nonparametric Scan Statistic (NPSS)

Have: $S \subseteq \{A_1 \times \ldots \times A_M\}$
$= \{s_1 \times \ldots \times s_M\}$

Select: $F(S) = max_\alpha\ \phi(\alpha, N_\alpha(S), N(S))$

Want: $max_S\ F(S)$

Assume: $\phi \uparrow$ w.r.t N_α
$\phi \downarrow$ w.r.t N and α
ϕ is convex

There Exist: $G(a_i) = \frac{1}{n(a_i)} \sum_{A_{-j}} I(p_{ij} \leq \alpha)$

Such That: $\max_{s_j \subseteq \{a_1, \ldots, a_t\}} F(s_j | A_{-j}) = \max_{i=1\ldots t} F\left(\left\{a_{(1)} \ldots a_{(t)}\right\} | A_{-j}\right)$

Only Consider: $\{$Black$\}$
$\{$Black, Hispanic$\}$
⋮
$\{$Black, Hispanic, Asian, …, White$\}$

I. Compute the statistical anomalousness of each treatment group subject

II. Discover subsets of attribute values that define the most anomalous outcomes

1. Maximize F(S) over all subsets of $s_1 \times \ldots \times s_M$
   - Naïve search is infeasible $O(2^{\Sigma|A_i|})$

Treatment Effects Subset Scan (TESS)

Nonparametric Scan Statistic (NPSS)

Have: $S \subseteq \{A_1 \times \ldots \times A_M\}$
$= \{s_1 \times \ldots \times s_M\}$

Select: $F(S) = max_\alpha\ \phi(\alpha, N_\alpha(S), N(S))$

Want: $max_S\ F(S)$

Assume: $\phi \uparrow$ w.r.t N_α
$\phi \downarrow$ w.r.t N and α
ϕ is convex

There Exist: $G(a_i) = \frac{1}{n(a_i)} \sum_{A_{-j}} I(p_{ij} \leq \alpha)$

Such That: $\max_{s_j \subseteq \{a_1, \ldots, a_t\}} F(s_j \mid A_{-j}) = \max_{i=1\ldots t} F\left(\{a_{(1)} \ldots a_{(t)}\} \mid A_{-j}\right)$

Intuitively: $F(\{a_{(1)}, a_{(3)}\}) \leq F(\{a_{(1)}, a_{(2)}\})$
$G(a_{(3)}) \leq G(a_{(2)})$

I. Compute the statistical anomalousness of each treatment group subject

II. Discover subsets of attribute values that define the most anomalous outcomes

1. Maximize F(S) over all subsets of $s_1 \times \ldots \times s_M$
   - Naïve search is infeasible $O(2^{\Sigma|A_i|})$

Treatment Effects Subset Scan (TESS)

Nonparametric Scan Statistic (NPSS)

Have: $S \subseteq \{A_1 \times \ldots \times A_M\}$
$= \{s_1 \times \ldots \times s_M\}$

Select: $F(S) = max_\alpha\ \phi(\alpha, N_\alpha(S), N(S))$

Want: $max_S\ F(S)$

Assume: $\phi \uparrow$ w.r.t N_α
$\phi \downarrow$ w.r.t N and α
ϕ is convex

There Exist: $G(a_i) = \frac{1}{n(a_i)} \sum_{A_{-j}} I\left(p_{ij} \leq \alpha\right)$

Such That: $\max_{s_j \subseteq \{a_1, \ldots, a_t\}} F(s_j \mid A_{-j}) = \max_{i=1\ldots t} F\left(\left\{a_{(1)} \ldots a_{(t)}\right\} \mid A_{-j}\right)$

Intuitively: $F(\{a_{(1)}, a_{(3)}\}) \leq F(\{a_{(1)}, a_{(2)}\})$
$G(a_{(3)}) \leq G(a_{(2)})$

I. Compute the statistical anomalousness of each treatment group subject

II. Discover subsets of attribute values that define the most anomalous outcomes

1. Maximize F(S) over all subsets of $s_1 \times \ldots \times s_M$
   - Naïve search is infeasible $O(2^{\Sigma|A_i|})$

Higher Criticism:

$$F(S) = \max_\alpha \frac{N_\alpha - N\alpha}{\sqrt{N\alpha(1-\alpha)}}$$

Treatment Effects Subset Scan (TESS)

Nonparametric Scan Statistic (NPSS)

Have: $S \subseteq \{A_1 \times \ldots \times A_M\}$
$= \{s_1 \times \ldots \times s_M\}$

Select: $F(S) = max_\alpha\, \phi(\alpha, N_\alpha(S), N(S))$

Want: $max_S\, F(S)$

Assume: $\phi \uparrow$ w.r.t N_α
$\phi \downarrow$ w.r.t N and α
ϕ is convex

There Exist: $G(a_i) = \frac{1}{n(a_i)} \sum_{A_{-j}} I\left(p_{ij} \leq \alpha\right)$

Such That: $\max_{s_j \subseteq \{a_1, \ldots, a_t\}} F(s_j \mid A_{-j}) = \max_{i=1\ldots t} F\left(\left\{a_{(1)} \ldots a_{(t)}\right\} \mid A_{-j}\right)$

Intuitively: $F(\{a_{(1)}, a_{(3)}\}) \leq F(\{a_{(1)}, a_{(2)}\})$
$G(a_{(3)}) \leq G(a_{(2)})$

I. Compute the statistical anomalousness of each treatment group subject

II. Discover subsets of attribute values that define the most anomalous outcomes

1. Maximize F(S) over all subsets of $s_1 \times \ldots \times s_M$
   - NPSS over an attribute in O(t log t)

Higher Criticism:

$$F(S) = \max_\alpha \frac{N_\alpha - N\alpha}{\sqrt{N\alpha(1-\alpha)}}$$

TESS Search Procedure

	Male	Female
Black	P^{BM}	P^{BF}
White	P^{WM}	P^{WF}
Hispanic	P^{HM}	P^{HF}
Asian	P^{AM}	P^{AF}
Native American	P^{NM}	P^{NF}
Other	P^{OM}	P^{OF}

Treatment Group

I. Compute the statistical anomalousness of each treatment group subject

1. Estimate Conditional Distribution Under H_0
2. Compute empirical p-values

II. Discover subsets of attribute values that define the most anomalous outcomes

1. Maximize F(S) over all subsets of $s_1 \times \ldots \times s_M$
   - NPSS over an attribute in O(t log t)

TESS Search Procedure

	Male	Female
Black	P^{BM}	P^{BF}
White	P^{WM}	P^{WF}
Hispanic	P^{HM}	P^{HF}
Asian	P^{AM}	P^{AF}
Native American	P^{NM}	P^{NF}
Other	P^{OM}	P^{OF}

(Score = 7.5)

I. Compute the statistical anomalousness of each treatment group subject

1. Estimate Conditional Distribution Under H_0
2. Compute empirical p-values

II. Discover subsets of attribute values that define the most anomalous outcomes

1. Maximize F(S) over all subsets of $s_1 \times \ldots \times s_M$
   - NPSS over an attribute in O(t log t)

TESS Search Procedure

	Male	Female
Black	P^{BM}	P^{BF}
White	P^{WM}	P^{WF}
Hispanic	P^{HM}	P^{HF}
Asian	P^{AM}	P^{AF}
Native American	P^{NM}	P^{NF}
Other	P^{OM}	P^{OF}

(Score = 8.1)

I. Compute the statistical anomalousness of each treatment group subject

1. Estimate Conditional Distribution Under H_0
2. Compute empirical p-values

II. Discover subsets of attribute values that define the most anomalous outcomes

1. Maximize F(S) over all subsets of $s_1 \times \ldots \times s_M$
   - NPSS over an attribute in O(t log t)

TESS Search Procedure

	Male	Female
Black	P^{BM}	P^{BF}
White	P^{WM}	P^{WF}
Hispanic	P^{HM}	P^{HF}
Asian	P^{AM}	P^{AF}
Native American	P^{NM}	P^{NF}
Other	P^{OM}	P^{OF}

(Score = 9.3)

I. Compute the statistical anomalousness of each treatment group subject

1. Estimate Conditional Distribution Under H_0
2. Compute empirical p-values

II. Discover subsets of attribute values that define the most anomalous outcomes

1. Maximize F(S) over all subsets of $s_1 \times \ldots \times s_M$
   - NPSS over an attribute in O(t log t)

TESS Search Procedure

	Male	Female
Black	P^{BM}	P^{BF}
White	P^{WM}	P^{WF}
Hispanic	P^{HM}	P^{HF}
Asian	P^{AM}	P^{AF}
Native American	P^{NM}	P^{NF}
Other	P^{OM}	P^{OF}

(Score = 9.3)

I. Compute the statistical anomalousness of each treatment group subject

1. Estimate Conditional Distribution Under H_0
2. Compute empirical p-values

II. Discover subsets of attribute values that define the most anomalous outcomes

1. Maximize F(S) over all subsets of $s_1 \times \ldots \times s_M$
   - NPSS over an attribute in O(t log t)

<u>Significance of our subpopulation</u>
Compare subpopulation score to maximum scores of simulated datasets under H_0

Agenda

- Introduction and Motivation
- Machine Learning's Contributions (and Limits)
- How Anomaly Detection Can Help
- Treatment Effect Subset Scan
  - Algorithm
  - Statistical Properties
- Results
- Conclusions

TESS's Statistical Properties

- False Positive under H_0 **(Theorem 2)**

$$P_{H_0}\left(\max_S F(S) > h(M)\right) \to 0$$

- Power under H_1 **(Theorem 3)**

$$P_{H_1}\left(\max_S F(S) > h(M)\right) \to 1$$

- Exactness of Detected Subpopulation
  - Sufficiently Homogenous **(Theorem 4)**

    $$S^* \supseteq S^T$$

  - Sufficiently Strong **(Theorem 5)**

    $$S^* \subseteq S^T$$

Agenda

- Introduction and Motivation
- Machine Learning's Contributions (and Limits)
- How Anomaly Detection Can Help
- Treatment Effect Subset Scan
  - Algorithm
  - Statistical Properties
- **Results**
- Conclusions

Tennessee Star Analysis (1985)

- Effect of classrooms size on achievement (test scores)

- 4 year panel (kindergarten to 3rd grade)

- 6,500 students, 330 classrooms, 80 schools
  - Total of over 11,000 records

- Treatment Conditions (randomized within school)
  - Regular Size Class (20-25 students)
  - Regular Size + Aide Class (20-25 students)
  - Small Size Class (13-17 students)

Tennessee Star Analysis (1985)

read	math	gender	ethnicity	lunch	grade	school	experience	degree	tethnicity	schoolid
439	463	male	afam	free	kindergarten	inner-city	0	bachelor	cauc	19
448	559	male	cauc	non-free	kindergarten	rural	16	bachelor	cauc	69
431	454	male	cauc	free	kindergarten	rural	8	bachelor	cauc	5
395	423	female	afam	free	kindergarten	inner-city	17	master	cauc	16
451	500	female	cauc	non-free	kindergarten	rural	3	bachelor	afam	56
430	473	male	cauc	non-free	kindergarten	rural	13	master	cauc	38
437	468	male	cauc	non-free	kindergarten	rural	6	master	cauc	69
490	528	male	cauc	non-free	kindergarten	suburban	18	bachelor	cauc	52
439	484	male	cauc	non-free	kindergarten	suburban	13	master	cauc	54
424	459	female	cauc	free	kindergarten	rural	12	bachelor	cauc	12
437	528	female	afam	free	kindergarten	suburban	1	bachelor	afam	21
424	559	male	cauc	free	kindergarten	rural	13	bachelor	cauc	79
431	454	male	cauc	non-free	kindergarten	rural	13	master	cauc	8
451	473	male	cauc	non-free	kindergarten	rural	3	bachelor	cauc	66
421	459	female	afam	free	kindergarten	inner-city	11	bachelor	cauc	31

Tennessee Star Analysis

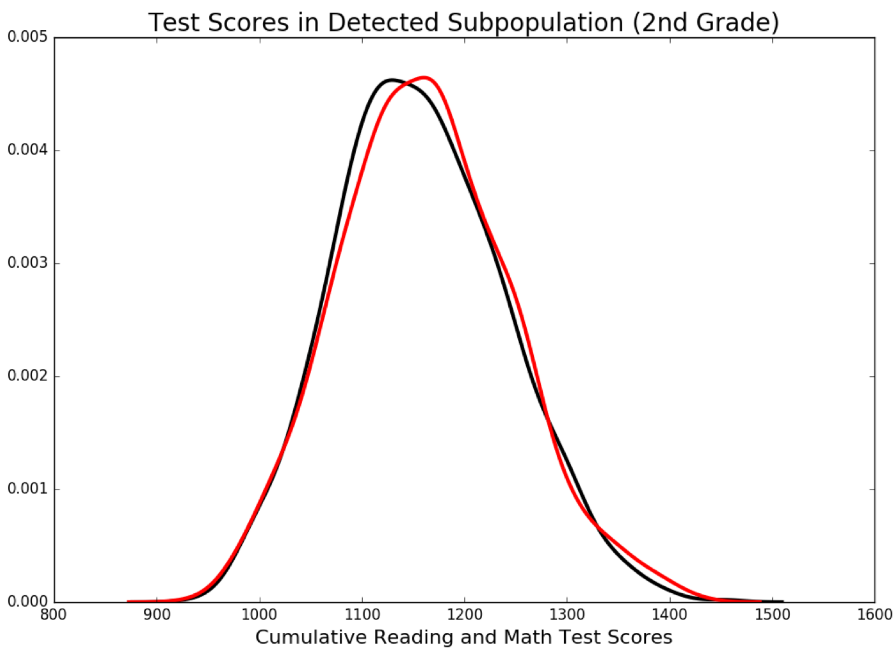
	(1)	(2)
Treatment	3.4791	-0.2909
	(2.547)	(2.277)
Sample	All 2nd Grade	All 3rd Grade
R-squared	0.000	0.000
Observations	4263	4063

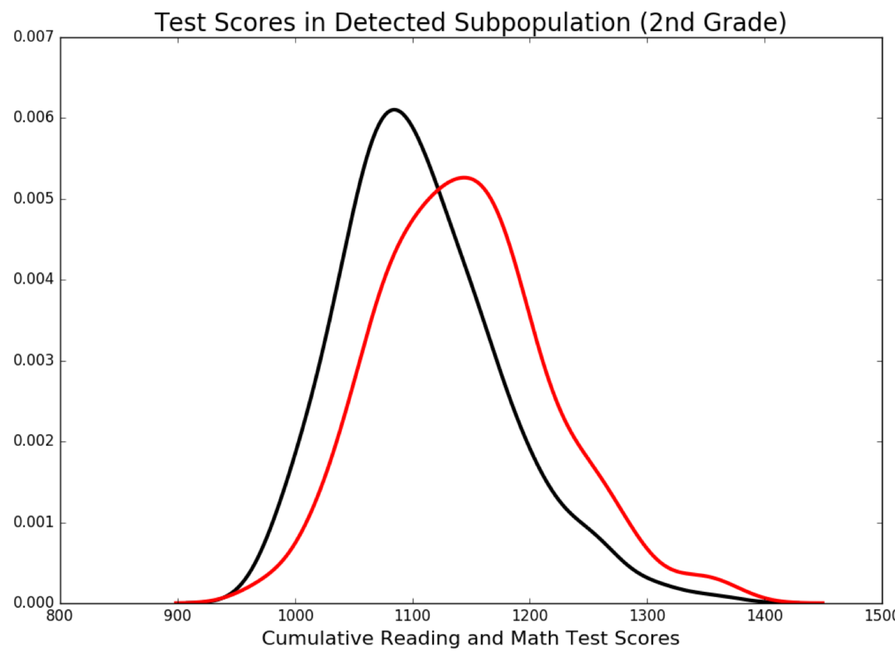
Notes: All estimates are from OLS models.
Standard errors are in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

Tennessee Star Analysis (1985)

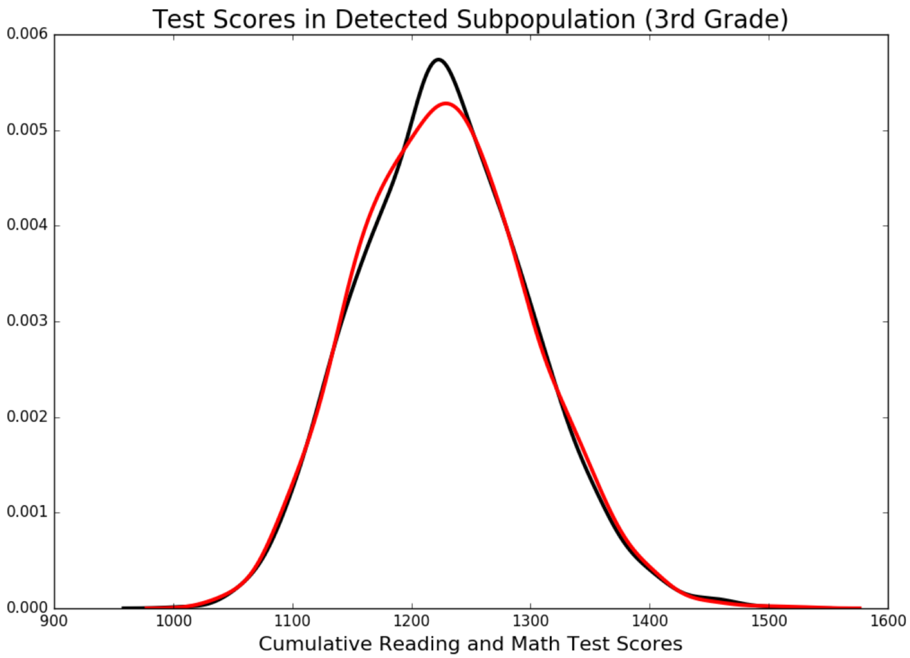
- Detected Subpopulation
  - grade:
    - 2nd or 3rd
  - school:
    - inner-city or urban
  - experience:
    - [10, infinity)
  - other features do not have differential effects

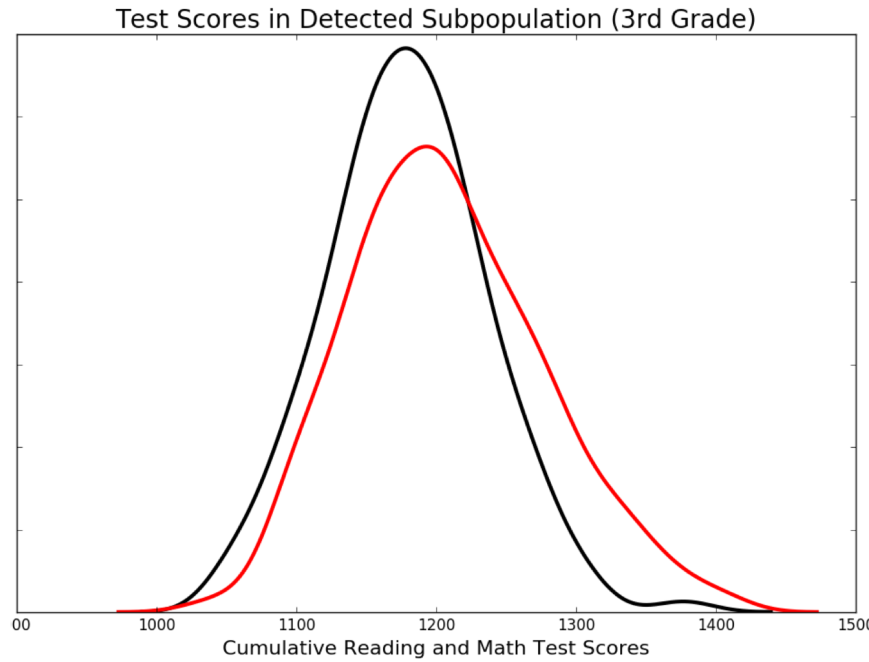
Tennessee Star Analysis





Tennessee Star Analysis





Tennessee Star Analysis

	(1)	(2)	(3)
Treatment	3.4791	36.066***	1.309
	(2.547)	(6.055)	(2.772)
Sample	All 2^{nd} Grade	Detected Group (2^{nd} Grade)	Undetected Group (2^{nd} Grade)
P-value	0.172	<0.001	0.0637
Observations	4263	620	3643

Notes: All estimates are from OLS models.
Standard errors are in parentheses.
*** p<0.001, ** p<0.05, * p<0.1

Tennessee Star Analysis

	(1)	(2)	(3)
Treatment	-0.291	18.703***	0.1
	(2.277)	(5.18)	(2.478)
Sample	All 3rd Grade	Detected Group (3rd Grade)	Undetected Group (3rd Grade)
P-value	0.898	<0.001	0.968
Observations	4063	706	3357

Notes: All estimates are from OLS models.
Standard errors are in parentheses.
*** p<0.001, ** p<0.05, * p<0.1

Conclusion

- Discovering subpopulations with significant treatment effects can be paramount
- Machine Learning can flexibly estimate effects but it limited when goal is to identify subpopulations with large effects
- Anomalous Pattern Detection paradigm offers overcome some abilities to overcome these limitations
  - Maintain high power to detect by searching over and combining signal across various subpopulations