

# Efficient Identification of Heterogeneous Treatment Effects via Anomalous Pattern Detection

Edward McFowland III

[emcfowla@umn.edu](mailto:emcfowla@umn.edu)

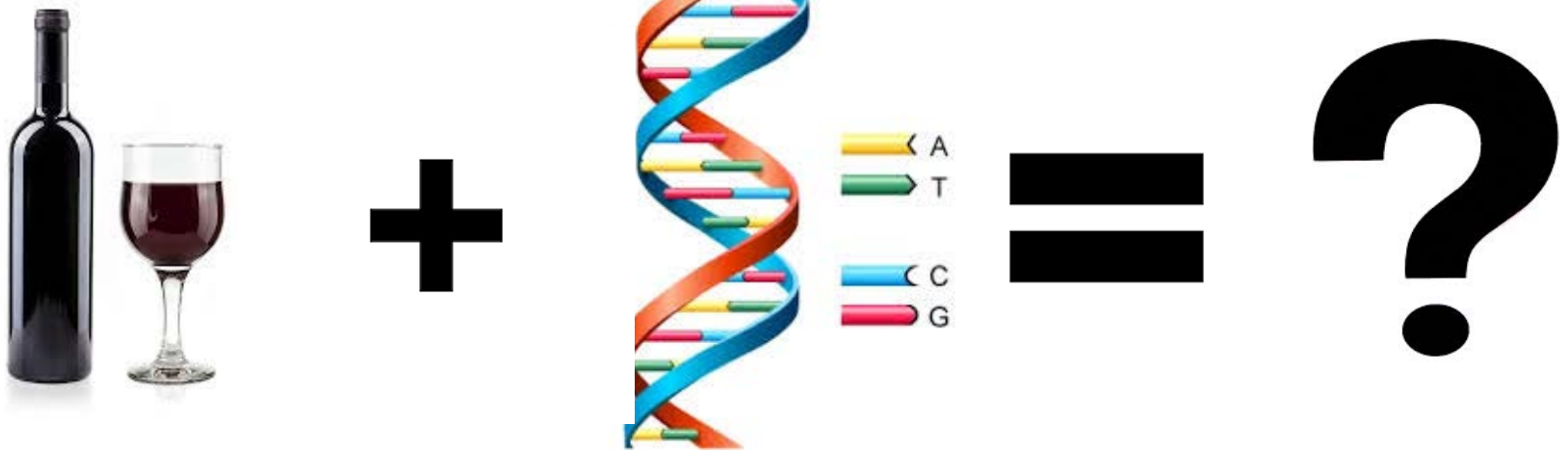
Joint work with (Sriram Somanchi & Daniel B. Neill)



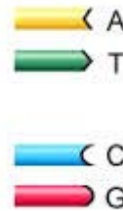
UNIVERSITY OF MINNESOTA

Driven to Discover<sup>SM</sup>

# Treatment Effects



# Treatment Effects



# Treatment Effect Heterogeneity


Control Group      Treatment Group


	Control Group	Treatment Group
Age = Young	YC YC YC	YT YT YT
Age = Mid	YC YC YC	YT YT YT
Age = Old	YC YC YC	YT YT YT


# Treatment Effect Heterogeneity


Control Group      Treatment Group

Age = Young	YC	YT
	YC	YT
	YC	YT
Age = Mid	YC	YT
	YC	YT
	YC	YT
Age = Old	YC	YT
	YC	YT
	YC	YT

 (Positive Effect)

 (Negative Effect)

 (No Effect)


 (No Effect)


- Positive and negative effects can cancel


# Treatment Effect Heterogeneity


Control Group      Treatment Group

	Control Group	Treatment Group
Age = Young	YC YC YC	YT YT YT
Age = Mid	YC YC YC	YT YT YT
Age = Old	YC YC YC	YT YT YT

 (Positive Effect)

 (No Effect)





 (No Effect)

 (No Effect)

- Positive and negative effects can cancel





- True effect can be masked

# Treatment Effect Heterogeneity

	<u>Control Group</u>	<u>Treatment Group</u>	
Age = Young	YC YC YC	YT YT YT	 ( Very Positive Effect )
Age = Mid	YC YC YC	YT YT YT	 ( No Effect )
Age = Old	YC YC YC	YT YT YT	 ( No Effect )
			 ( Slight Effect )

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

# Treatment Effect Heterogeneity

	Control Group	Treatment Group	
Age = Young	YC YC YC	YT YT YT	 ( Very Positive Effect )
Age = Mid	YC YC YC	YT YT YT	 ( No Effect )
Age = Old	YC YC YC	YT YT YT	 ( No Effect )
			 ( Slight Effect )

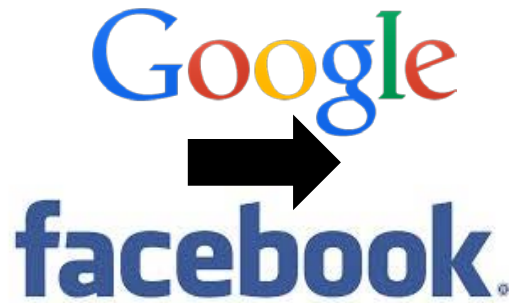
- Positive and negative effects can cancel
- True effect can be masked
  - Ex: FDA Approved BiDil Drug
- Effects could really be driven by a subpopulation
  - Ex: Perry Preschool



# Treatment Effect Heterogeneity (Big Data)

Control      Treatment  
Group        Group

YC	YT
YC	YT
YC	YT
YC	YT
YC	YT
YC	YT
YC	YT
YC	YT
YC	YT
YC	YT



YC	YC	YC	YC	YC	YC	YC	YC
YT	YT	YT	YT	YT	YT	YT	YT
YC	YC	YC	YC	YC	YC	YC	YC
YT	YT	YT	YT	YT	YT	YT	YT
YC	YC	YC	YC	YC	YC	YC	YC
YT	YT	YT	YT	YT	YT	YT	YT
YC	YC	YC	YC	YC	YC	YC	YC
YT	YT	YT	YT	YT	YT	YT	YT
YC	YC	YC	YC	YC	YC	YC	YC
YT	YT	YT	YT	YT	YT	YT	YT

⋮

YC	YC	YC	YC	YC	YC	YC	YC
YT	YT	YT	YT	YT	YT	YT	YT

# Machine Learning's Contributions

- Regression Methods
  - OLS and Regularized Regression (e.g., LASSO)\*
  - Imai and Ratkovic (2013)
- Single Tree Methods
  - Su et al (2009)
  - Imai and Strauss (2011)
  - Athey and Imbens (2015)\*
- Ensemble Methods
  - Green and Kern (2012)
  - Wager and Athey (2012)\*

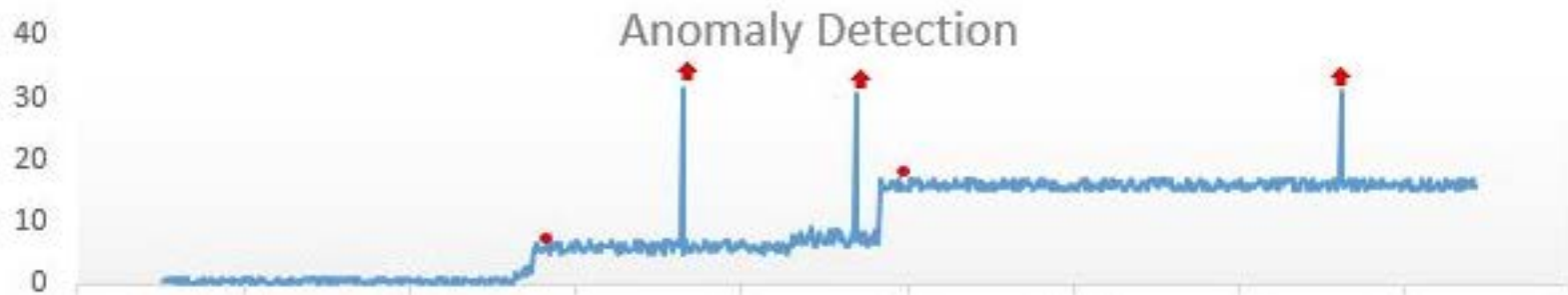
\* provide asym confidence intervals, for inference of effects significance

# 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

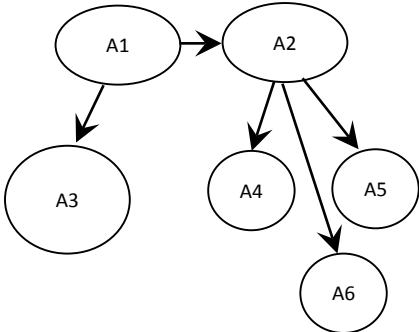
# Anomaly Detection Paradigm

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



# Anomalous Pattern Detection Procedure

Normal Activity ( $M_0$ )



PORT	ISPORT	COUNTRY	BLDG	VESSEL	SHIPPERNAME	FINAME	COMMODITY	SIZE	MTONS	VALUE
YOKOHAMA	SEATTLE	JAPAN	OSCO	LANG YUN HE	AMERICAN TRU NET	SUPRESTR	NET BUFTY RACK	0	56	2979
YOKOHAMA	SEATTLE	JAPAN	OSCO	LANG YUN HE	ORDER	ORDER	USED TREE	2	1343	847
YOKOHAMA	SEATTLE	JAPAN	OSCO	LANG YUN HE	ORDER	ORDER	USED TREE	2	1343	847
YOKOHAMA	SEATTLE	JAPAN	OSCO	LANG YUN HE	AMERICAN TRU NET	SUPRESTR	NET ORDER CODE PARTY	1	1700	2819
YOKOHAMA	SEATTLE	JAPAN	OSCO	LANG YUN HE	NEW HAVR	TRANSPORT	JIT PAVES F MODEL SR	3	357	8816
YOKOHAMA	SEATTLE	JAPAN	OSCO	LANG YUN HE	NEW HAVR	TRANSPORT	JIT PAVES F MODEL SR	3	357	8816
YOKOHAMA	SEATTLE	JAPAN	OSCO	LANG YUN HE	NEW HAVR	TRANSPORT	JIT PAVES F MODEL SR	3	357	8816
YOKOHAMA	SEATTLE	JAPAN	OSCO	LANG YUN HE	ORDER	ORDER	USED TREES	2	1343	847
YOKOHAMA	SEATTLE	JAPAN	OSCO	LANG YUN HE	CHINA OCEAN	SPGS	CHINA OCEAN CONTAINERS	0	0	0
YOKOHAMA	SEATTLE	JAPAN	OSCO	LANG YUN HE	CHINA OCEAN	SPGS	CHINA OCEAN CONTAINERS	0	0	0

Test Data

Detect Anomalous Pattern Given  $M_0$

Novel Pattern

YOKOHAMA	SEATTLE	JAPAN	OSCO	LANG YUN HE	ORDER	ORDER	USED TREE	2	1343	847
YOKOHAMA	SEATTLE	JAPAN	OSCO	LANG YUN HE	ORDER	ORDER	USED TREE	2	1343	847
YOKOHAMA	SEATTLE	JAPAN	OSCO	LANG YUN HE	AMERICAN TRU NET	SUPRESTR	NET ORDER CODE PARTY	1	1700	2819
YOKOHAMA	SEATTLE	JAPAN	OSCO	LANG YUN HE	NEW HAVR	TRANSPORT	JIT PAVES F MODEL SR	3	357	8816

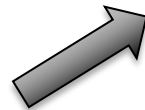
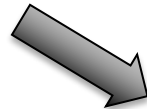
# HTE Pattern Detection

Control Group

LINE	DATE	TIME	FROM	TO	STATUS	REASON
1001	1/1/2000	10:00	SEA	SEA	OK	
1002	1/1/2000	10:00	SEA	SEA	OK	
1003	1/1/2000	10:00	SEA	SEA	OK	
1004	1/1/2000	10:00	SEA	SEA	OK	
1005	1/1/2000	10:00	SEA	SEA	OK	
1006	1/1/2000	10:00	SEA	SEA	OK	
1007	1/1/2000	10:00	SEA	SEA	OK	
1008	1/1/2000	10:00	SEA	SEA	OK	
1009	1/1/2000	10:00	SEA	SEA	OK	
1010	1/1/2000	10:00	SEA	SEA	OK	

Treatment Group

LINE	DATE	TIME	FROM	TO	STATUS	REASON
1001	1/1/2000	10:00	SEA	SEA	OK	
1002	1/1/2000	10:00	SEA	SEA	OK	
1003	1/1/2000	10:00	SEA	SEA	OK	
1004	1/1/2000	10:00	SEA	SEA	OK	
1005	1/1/2000	10:00	SEA	SEA	OK	
1006	1/1/2000	10:00	SEA	SEA	OK	
1007	1/1/2000	10:00	SEA	SEA	OK	
1008	1/1/2000	10:00	SEA	SEA	OK	
1009	1/1/2000	10:00	SEA	SEA	OK	
1010	1/1/2000	10:00	SEA	SEA	OK	



Detect Anomalous Pattern Given  $M_0$



Novel Pattern

1001	SEA	SEA	OK		
1002	SEA	SEA	OK		
1003	SEA	SEA	OK		
1004	SEA	SEA	OK		
1005	SEA	SEA	OK		
1006	SEA	SEA	OK		
1007	SEA	SEA	OK		
1008	SEA	SEA	OK		
1009	SEA	SEA	OK		
1010	SEA	SEA	OK		

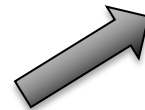
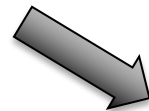
# HTE Pattern Detection

Control  
Group

USER	DATE	TIME	LOCATION	EVENT	STATUS	REASON
U001	1/1/2020	10:00	SEA-TOLSON	ENTER	OK	
U002	1/1/2020	10:05	SEA-TOLSON	ENTER	OK	
U003	1/1/2020	10:10	SEA-TOLSON	ENTER	OK	
U004	1/1/2020	10:15	SEA-TOLSON	ENTER	OK	
U005	1/1/2020	10:20	SEA-TOLSON	ENTER	OK	
U006	1/1/2020	10:25	SEA-TOLSON	ENTER	OK	
U007	1/1/2020	10:30	SEA-TOLSON	ENTER	OK	
U008	1/1/2020	10:35	SEA-TOLSON	ENTER	OK	
U009	1/1/2020	10:40	SEA-TOLSON	ENTER	OK	
U010	1/1/2020	10:45	SEA-TOLSON	ENTER	OK	

Treatment  
Group

USER	DATE	TIME	LOCATION	EVENT	STATUS	REASON
U001	1/1/2020	10:00	SEA-TOLSON	ENTER	OK	
U002	1/1/2020	10:05	SEA-TOLSON	ENTER	OK	
U003	1/1/2020	10:10	SEA-TOLSON	ENTER	OK	
U004	1/1/2020	10:15	SEA-TOLSON	ENTER	OK	
U005	1/1/2020	10:20	SEA-TOLSON	ENTER	OK	
U006	1/1/2020	10:25	SEA-TOLSON	ENTER	OK	
U007	1/1/2020	10:30	SEA-TOLSON	ENTER	OK	
U008	1/1/2020	10:35	SEA-TOLSON	ENTER	OK	
U009	1/1/2020	10:40	SEA-TOLSON	ENTER	OK	
U010	1/1/2020	10:45	SEA-TOLSON	ENTER	OK	



Detect Anomalous  
Subpopulation  
Given  
 $M_0$



Novel Pattern

U001	1/1/2020	10:00	SEA-TOLSON	ENTER	OK	
U002	1/1/2020	10:05	SEA-TOLSON	ENTER	OK	
U003	1/1/2020	10:10	SEA-TOLSON	ENTER	OK	
U004	1/1/2020	10:15	SEA-TOLSON	ENTER	OK	
U005	1/1/2020	10:20	SEA-TOLSON	ENTER	OK	
U006	1/1/2020	10:25	SEA-TOLSON	ENTER	OK	
U007	1/1/2020	10:30	SEA-TOLSON	ENTER	OK	
U008	1/1/2020	10:35	SEA-TOLSON	ENTER	OK	
U009	1/1/2020	10:40	SEA-TOLSON	ENTER	OK	
U010	1/1/2020	10:45	SEA-TOLSON	ENTER	OK	

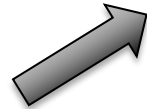
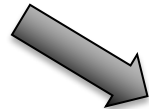
# HTE Pattern Detection

Control  
Group

USER	IP	OS	APP	TIME	STATUS
1001	192.168.1.1	Windows	Chrome	10:00	Success
1002	192.168.1.2	Windows	Chrome	10:01	Success
1003	192.168.1.3	Windows	Chrome	10:02	Success
1004	192.168.1.4	Windows	Chrome	10:03	Success
1005	192.168.1.5	Windows	Chrome	10:04	Success
1006	192.168.1.6	Windows	Chrome	10:05	Success
1007	192.168.1.7	Windows	Chrome	10:06	Success
1008	192.168.1.8	Windows	Chrome	10:07	Success
1009	192.168.1.9	Windows	Chrome	10:08	Success
1010	192.168.1.10	Windows	Chrome	10:09	Success

Treatment  
Group

USER	IP	OS	APP	TIME	STATUS
1011	192.168.1.11	Windows	Chrome	10:10	Success
1012	192.168.1.12	Windows	Chrome	10:11	Success
1013	192.168.1.13	Windows	Chrome	10:12	Success
1014	192.168.1.14	Windows	Chrome	10:13	Success
1015	192.168.1.15	Windows	Chrome	10:14	Success
1016	192.168.1.16	Windows	Chrome	10:15	Success
1017	192.168.1.17	Windows	Chrome	10:16	Success
1018	192.168.1.18	Windows	Chrome	10:17	Success
1019	192.168.1.19	Windows	Chrome	10:18	Success
1020	192.168.1.20	Windows	Chrome	10:19	Success



Detect Anomalous  
Subpopulation  
Given  
 $M_0$



Subpopulation

1021	192.168.1.21	Windows	Chrome	10:20	Success
1022	192.168.1.22	Windows	Chrome	10:21	Success
1023	192.168.1.23	Windows	Chrome	10:22	Success
1024	192.168.1.24	Windows	Chrome	10:23	Success
1025	192.168.1.25	Windows	Chrome	10:24	Success



# 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

# 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

## The Optimization

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

# 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

## The Optimization

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

$$S = s_1 \times \dots \times s_M$$

# 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

## The Optimization

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

$$S = s_1 \times \dots \times s_M$$

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

# Treatment Effects Subset Scan (TESS)

Male      Female

Black	$\gamma^{BM}$	$\gamma^{BF}$
White	$\gamma^{WM}$	$\gamma^{WF}$
Hispanic	$\gamma^{HM}$	$\gamma^{HF}$
Asian	$\gamma^{AM}$	$\gamma^{AF}$
Native American	$\gamma^{NM}$	$\gamma^{NF}$
Other	$\gamma^{OM}$	$\gamma^{OF}$

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	$\gamma^{BM}$	$\gamma^{BF}$
White	$\gamma^{WM}$	$\gamma^{WF}$
Hispanic	$\gamma^{HM}$	$\gamma^{HF}$
Asian	$\gamma^{AM}$	$\gamma^{AF}$
Native American	$\gamma^{NM}$	$\gamma^{NF}$
Other	$\gamma^{OM}$	$\gamma^{OF}$
	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	$\gamma^{BM}$	$\gamma^{BF}$
White	$\gamma^{WM}$	$\gamma^{WF}$
Hispanic	$\gamma^{HM}$	$\gamma^{HF}$
Asian	$\gamma^{AM}$	$\gamma^{AF}$
Native American	$\gamma^{NM}$	$\gamma^{NF}$
Other	$\gamma^{OM}$	$\gamma^{OF}$
	Treatment Group	

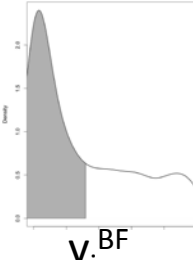
- I. Compute the statistical anomalousness of each treatment group subject
  1. Estimate Conditional Distribution Under  $H_0$
  2. Compute empirical p-values



# Treatment Effects Subset Scan (TESS)

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

Treatment Group



The graph shows a distribution curve with a shaded area under the curve, labeled with the symbol  $\gamma_i^{BF}$ . A blue arrow points from the 'Female' column of the table to the graph.

- 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.  $P_{ij} \sim \text{Uniform}[0,1]$  under  $H_0$
    - iii. For any  $N$  p-values, we expect  $N \cdot \alpha$  to be significant at level  $\alpha$

# 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
  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  $H_0$
    - iii. For any  $N$  p-values, we expect  $N*\alpha$  to be significant at level  $\alpha$

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  $S_1 \times \dots \times S_M$ 
    - Naïve search is infeasible  $O(2^{\sum |A_i|})$

# Treatment Effects Subset Scan (TESS)

## Nonparametric Scan Statistic (NPSS)

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

Select:  $F(S)$

Want:  $\max_S F(S)$

- 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  $s_1 \times \dots \times s_M$ 
    - Naïve search is infeasible  $O(2^{\sum |A_i|})$



# Treatment Effects Subset Scan (TESS)

## Nonparametric Scan Statistic (NPSS)

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

Select:  $F(S)$

Want:  $\max_S F(S)$

There Exist:  $G(a_i)$

- 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  $s_1 \times \dots \times s_M$ 
    - Naïve search is infeasible  $O(2^{\sum |A_i|})$

# Treatment Effects Subset Scan (TESS)

## Nonparametric Scan Statistic (NPSS)

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

Select:  $F(S)$

Want:  $\max_S F(S)$

There Exist:  $G(a_i)$

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

Only Consider: {Black}

{Black, Hispanic}

{Black, Hispanic, Asian}

⋮

{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  $s_1 \times \dots \times s_M$ 
    - Naïve search is infeasible  $O(2^{\sum |A_i|})$

# Treatment Effects Subset Scan (TESS)

## Nonparametric Scan Statistic (NPSS)

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

Select:  $F(S)$

Want:  $\max_S F(S)$

There Exist:  $G(a_i)$

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

Only Consider: {Black}

{Black, Hispanic}

{Black, Hispanic, Asian}

⋮


{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  $s_1 \times \dots \times s_M$ 
    - 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	$p^{OM}$	$p^{OF}$

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  $S_1 \times \dots \times S_M$ 
    - 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  $S_1 \times \dots \times S_M$ 
    - 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  $S_1 \times \dots \times S_M$ 
    - 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  $S_1 \times \dots \times S_M$ 
    - 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  $S_1 \times \dots \times S_M$ 
    - NPSS over an attribute in  $O(t \log t)$

Significance of our subpopulation  
Compare subpopulation score to maximum scores of simulated datasets under  $H_0$



# Tennessee Star Analysis (1985)

- Effect of classrooms size on achievement (test scores)
- 4 year panel (kindergarten to 3<sup>rd</sup> 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 2 <sup>nd</sup> Grade	All 3 <sup>rd</sup> Grade
R-squared	0.000	0.000
Observations	4263	4063

Notes: All estimates are from OLS models.

Standard errors are in parentheses.

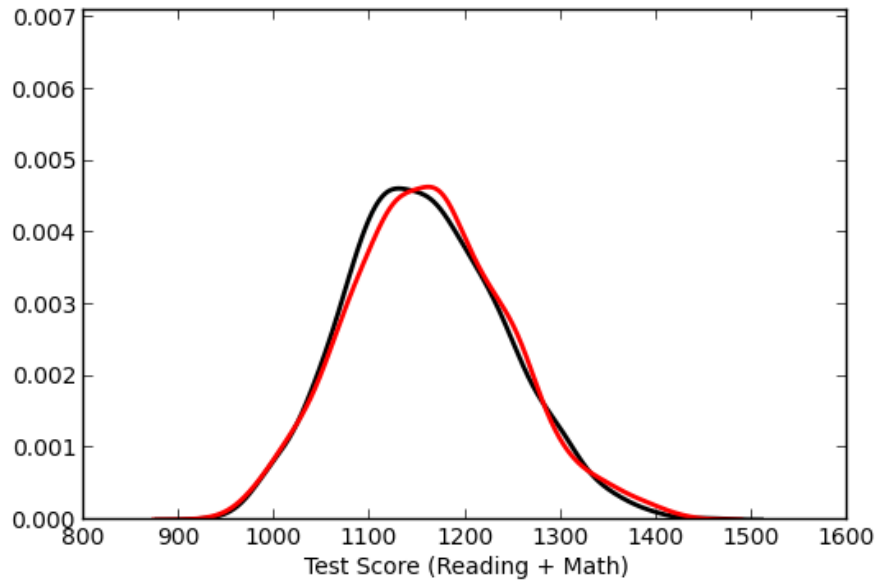
\*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.1$

# Tennessee Star Analysis (1985)

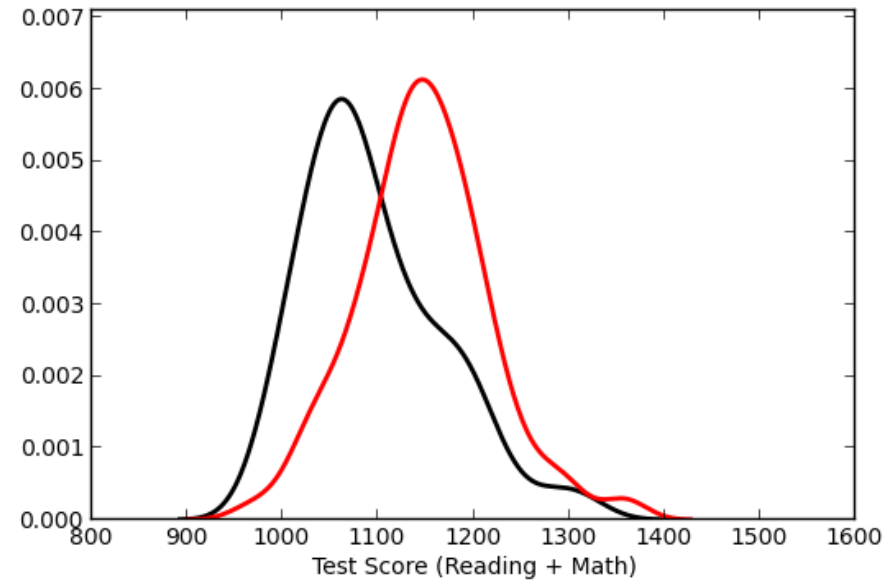
- Detected Subpopulation
  - grade:
    - 2nd or 3rd
  - school:
    - inner-city or urban
  - experience:
    - [10, infinity)
  - other features are considered irrelevant

# Tennessee Star Analysis

Outcome Distribution (2nd Grade)  
Treatment=2048 vs Control=2215

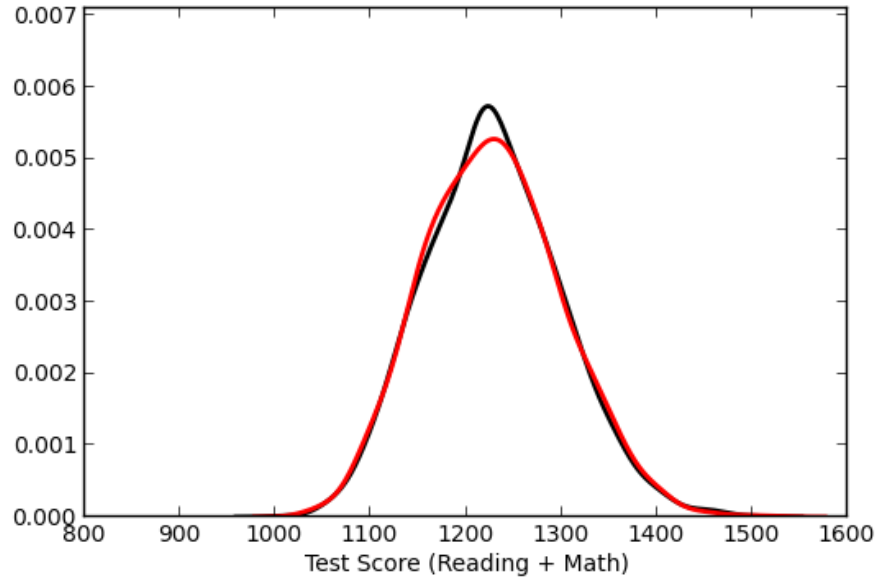


Subpopulation Outcome Distribution (2nd Grade)  
Treatment=104 vs Control=183

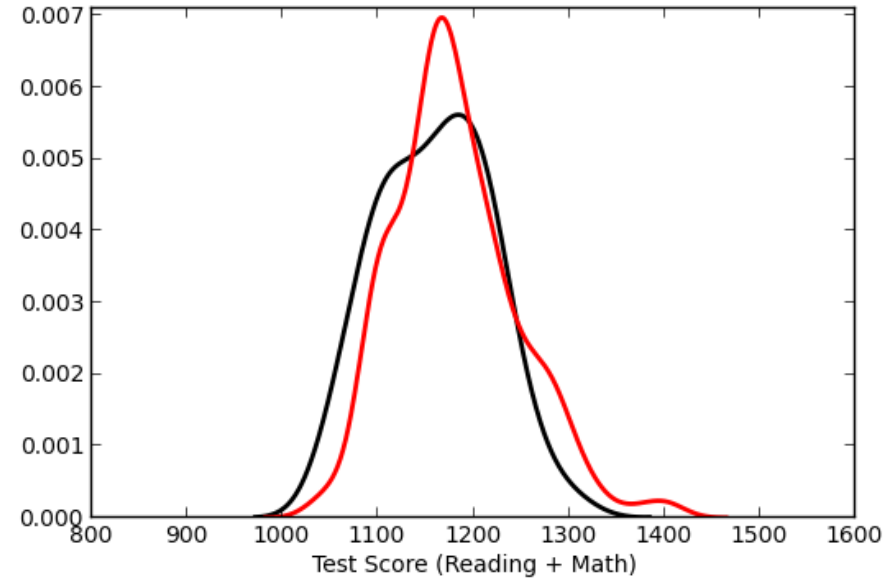


# Tennessee Star Analysis

Outcome Distribution (3rd Grade)  
Treatment=1807 vs Control=2256



Subpopulation Outcome Distribution (3rd Grade)  
Treatment=63 vs Control=195



# Tennessee Star Analysis

	(1)	(2)	(3)
Treatment	3.4791	51.2497***	1.4532
	(2.547)	(8.727)	(2.639)
Sample	All 2 <sup>nd</sup> Grade	Detected Subpopulation	Undetected Subpopulation
R-squared	0.001<	0.108	0.001<
Observations	4263	287	3976

Notes: All estimates are from OLS models.

Standard errors are in parentheses.

\*\*\* p<0.01, \*\* p<0.05, \* p<0.1

# Tennessee Star Analysis

	(1)	(2)	(3)
Treatment	-0.2909	21.822**	1.6851
	(2.277)	(9.154)	(2.318)
Sample	All 3 <sup>rd</sup> Grade	Detected Group (3 <sup>rd</sup> Grade)	Undetected Group (3 <sup>rd</sup> Grade)
R-squared	0.001<	0.022	0.001<
Observations	4063	258	3805

Notes: All estimates are from OLS models.

Standard errors are in parentheses.

\*\*\* p<0.01, \*\* p<0.05, \* p<0.1



# Conclusion

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

# References

- Athey, S., & Imbens, G. W. (2015). Machine Learning Methods for Estimating Heterogeneous Causal Effects. arXiv.org.
- Green, D. P., & Kern, H. L. (2012). Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees. *Public Opinion Quarterly*, 76(3), 491–511.
- Imai, K., & Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1), 443–470.
- Imai, K., & Strauss, A. (2011). Estimation of Heterogeneous Treatment Effects from Randomized Experiments, with Application to the Optimal Planning of the Get-Out-the-Vote Campaign. *Political Analysis*, 19(1), 1–19.
- McFowland, E., III, Speakman, S. D., & Neill, D. B. (2013). Fast generalized subset scan for anomalous pattern detection. *The Journal of Machine Learning Research*, 14(1), 1533–1561.
- Mehlig, Kirsten et al. CETP TaqIB genotype modifies the association between alcohol and coronary heart disease: The INTERGENE case-control study. *Alcohol*, 48(7), 695 - 700.
- Su, X., Tsai, C.-L., Wang, H., Nickerson, D. M., & Li, B. (2009). Subgroup Analysis via Recursive Partitioning. *The Journal of Machine Learning Research*, 10, 141–158.
- Wager, S., & Athey, S. (2015). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. arXiv.org.