Virtual Roundtable: Translating Data Science
Bias in Data Sets and Analytics

February 20, 2019

Ensuring the proper curation and integrity of data from the earliest stages of collection is absolutely essential. For the quality of downstream results to be trusted, bias must be minimized and accounted for through thoughtful quality control methodologies. As well, thoughtful considerations need to be made as to what types of data should be used in an analytics process, as to not create preemptive bias for a certain result.

Join us as we explore how to best manage these critical concerns.

Hareesh Chandrupatla
Chief Data Scientist
Anjin Analytics

Yacine Abdous
Vice President of Data Services
Cruz Street

Scott Lett, Ph.D.
Consulting Data Detective
Holistic Mathematics

Jim Duarte
Data Scientist/Consultant Advisor
Cruz Street
Westchester Biotech Project
Research Community Resources

-a borderless initiative mapping the future for regional and international collaboration

On Twitter we’re @WestchesterBio

Co-Founders:

Michael Welling, Chair
Partner, Meridian Risk Management

Joanne Gere, Executive Director
Westchester Biotech Project
333 Mamaroneck Avenue, #340
White Plains, New York 10605
917-719-2226
www.WestchesterBiotechProject.org
Thank You to our Community Partners, Alliance Partners, and Participants!
Virtual Roundtable: **Translating Data Science**

**Bias in Data Sets and Analytics**

**February 20, 2019**

Co-Chairs:

- **Paul Savage**  
  Healthcare MBA,  
  Hagan School of Business  
  Iona College

- **Antonio Biancardi**  
  CEO  
  Centrific Technology Partners

Steering Committee:

- **Teresita Abay-Krueger**, Owner/President, Tibi A Kogitate Consulting
- **Rene Baston**, Executive Director, Northeast Big Data Innovation
- **Hareesh Chandrupatla**, Chief Data Scientist, Anjin Analytics
- **Jesse Cruz**, President and Founder, Cruz Street
- **Karine Kleinhaus**, MD, MPH, Divisional Vice President, North America, Pluristem Therapeutics
- **Scott Lett, Ph.D.**, Consulting Data Detective, Holistic Mathematics
- **Angela Radcliffe**, R&D Practice Lead, Life Science, Capgemini Invent
- **Carla Romney, D.Sc.**, Adjunct Research Assistant Professor of Medical Sciences and Education, Boston University
- **Wade Trappe, Ph.D.**, Associate Director, WINLAB, Rutgers University
- **Marjorie Wilkie, Ph.D.**, Owner, Marjorie Wilke Consulting and Genetic Genealogy
Virtual Roundtable: **Translating Data Science**

**Bias in Data Sets and Analytics**

**February 20, 2019**

Hareesh Chandrupatla  
Chief Data Scientist  
Anjin Analytics

Scott Lett, Ph.D.  
Consulting Data Detective  
Holistic Mathematics

Yacine Abdous  
Vice President of Data Services  
Cruz Street

Jim Duarte  
Data Scientist/Consultant Advisor  
Cruz Street
Virtual Roundtable: Translating Data Science
Bias in Data Sets and Analytics

February 20, 2019

Ensuring the proper curation and integrity of data from the earliest stages of collection is absolutely essential. For the quality of downstream results to be trusted, bias must be minimized and accounted for through thoughtful quality control methodologies. As well, thoughtful considerations need to be made as to what types of data should be used in an analytics process, as to not create preemptive bias for a certain result.

Join us as we explore how to best manage these critical concerns.

Hareesh Chandrupatla
Chief Data Scientist
Anjin Analytics
Translating Data Science: Bias in Data Sets & Analytics

Westchester Biotech Project
February 20, 2019

ANJIN ANALYTICS
Advanced Analytics for Healthcare & Life Science

Hareesh Chandrupatla
hareesh@anjinanalytics.com
In The News

Amazon Recruitment

• “Amazon scraps secret AI recruiting tool that showed bias against women”
• “Amazon’s system taught itself that male candidates were preferable”
• Article
  • https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

• Bias
  • Selection
  • Confirmation
In The News

Machine Bias in Prison Sentencing

• COMPAS: Software provides risk scores used by judges for sentencing
• Higher Rate of False Positives for African Americans
• 137 questions: doesn’t ask for race directly but questions can be correlated to race including salary, family situation, location
• Article:
  • https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
• Bias
  • Selection
  • Confirmation
Use Case: Wound Care Clinical Decision Support

- Case: Wound care for patients with chronic wounds
- Goal: Automated Detection of Perimeter & Area of Wound from Images
- Data Set: 20000+ images of chronic wounds
  - 70% Caucasian
  - 20% African American
  - 10% South Asian
- Algorithm
  - Deep learning using CNN (Convolutional neural networks)
    - Detects wound boundary based on differences in pigmentation between healthy skin and wound
- Approach
  - Differences in skin pigmentation in population requires independent development of algorithm for each population
- Bias Issues
  - Selection bias leads to higher performing predictions for Caucasian vs other populations
Ensuring the proper curation and integrity of data from the earliest stages of collection is absolutely essential. For the quality of downstream results to be trusted, bias must be minimized and accounted for through thoughtful quality control methodologies. As well, thoughtful considerations need to be made as to what types of data should be used in an analytics process, as to not create preemptive bias for a certain result.

Join us as we explore how to best manage these critical concerns.
Bias In Machine Learning

Sources and Mitigation of Bias in Medical Applications
Scott Lett
- PhD Applied Math

Modeling, simulation, machine learning, data science in
- Space
  - Spacecraft design, remote sensing, mission design
- Underground
  - Water and petroleum reservoir modeling
- Cloud computing
  - Oracle Cloud, entertainment industry
- Biomedical Applications
  - Diagnostics, drug safety, personalized cancer treatment
Bias

- Systematic Error
- Can result in undesirable errors in medical decisions.
- Presents risks to patients, insurance, health care/technology providers.
Traditional Types of Bias

- Sample Bias
- Prejudice Bias
- Measurement Bias
- Algorithm Bias
Sample Bias

- **Sample Size**
  - Budget limited clinical trials
  - One attempt to mitigate: meta-analysis
    - Pool data from multiple studies, improving results
      - “Osteopontin is a marker for cancer aggressiveness and patient survival,” Weber, Lett, Haubein

- **Publication Bias**
  - Studies with significant results 3 times more likely to be published.
  - Some techniques to detect and even mitigate.
Sample Bias

- Confounding Variables
  - Stirling’s Paradox

- Example: 1896 Hoffman published an exhaustive 330p statistical study that concluded the entire race of black people was uninsurable. He failed to account for poverty, injustice, etc.
  - Meta-analysis cannot overcome this statistical mistake.
Confounding Variables

- Example: cardiac drug safety.
  - Main variable: IKr sodium channel binding.
  - Confounding variables: Sex, health history, INa, IKs, Ica binding, genetic factors (race correlated), drug interactions, food interactions, HR.
Machine Learning to Infer Drug Effect on Confounding Channels

Input Data

Predicted EKG
Prejudice Bias

- Confirmation Bias
  - Desired (or predetermined result) confirmed by incorrect model. (See e.g. Hoffman 1896)

- Availability Bias
  - We make quick decisions based upon the immediately available information, especially if it confirms our bias.

- Possible main causes of reproducibility problem in science.
Measurement Bias

- Differences in modalities, protocols, produce different measurements in different studies.
- Challenge for meta-analysis, especially quantitative analysis.
  - In the Osteopontin work, we reduced quantitative data to categorical, and found that the groups with highest Osteopontin correlated with the most aggressive cancers and shortest survivals in almost all cancers.
Algorithm Bias

- Maximum Likelihood Underestimates Number of Cards in Hat

(Unbiased answer: 11)
Non-traditional Bias

- **Purpose Bias**
  - Monopoly pricing (instead of e.g. driving record) for unprivileged groups
    - *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy – O’Neil*
  - Fraud, e.g. targeting a susceptible group

- **Access Bias**
  - When life-changing decisions are made with black box models that you can’t cross examine, and the models put unprivileged groups at a disadvantage.

- **Model Bias**
  - When e.g. false positive rate is worse for unprivileged group

**Fairness**
AI Fairness 360

- Software framework for measuring and mitigating unfairness in AI models & data.
  - Open Source (from IBM) ([https://github.com/IBM/AIF360](https://github.com/IBM/AIF360))
  - >20 fairness metrics
  - >10 mitigations

- Model Training Data – new kinds
  - Attributes (used to predict outcome)
    - Protected vs Unprotected (partitions people into groups that should have parity)
    - Privileged Protected (groups that are historically at a systematic advantage)
  - Labels (outcomes)
    - Favorable/Unfavorable labels

- Fairness Metrics
  - Group fairness, individual fairness

- Mitigation
  - Preprocessing algorithms, in-processing algorithms, post-processing algorithms

- Explainers
  - Describes details or causes of fairness result

- Industry-specific guidance
Thoughts

- Guidance
- Tools
- Data – pooling, labeling, fairness assessment
- Funding
- Incentives
  - Publishing
  - Deploying to market
Virtual Roundtable: Translating Data Science
Bias in Data Sets and Analytics
February 20, 2019

Ensuring the proper curation and integrity of data from the earliest stages of collection is absolutely essential. For the quality of downstream results to be trusted, bias must be minimized and accounted for through thoughtful quality control methodologies. As well, thoughtful considerations need to be made as to what types of data should be used in an analytics process, as to not create preemptive bias for a certain result.

Join us as we explore how to best manage these critical concerns.

Yacine Abdous
Vice President of Data Services
Cruz Street

Jim Duarte
Data Scientist/Consultant Advisor
Cruz Street
Healthcare Data Quality Impact on Bias

Presented by Cruz Street & Expert Consultant Jim Duarte
## DOWNSTREAM CONSEQUENCES OF BAD DATA

<table>
<thead>
<tr>
<th>Tools For Improvement</th>
<th>Aspirational Analytics</th>
<th>Common Healthcare Gaps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Profiling</td>
<td>High Quality</td>
<td>Information Systems Governance</td>
</tr>
<tr>
<td>Discovery</td>
<td>Well Documented</td>
<td>Poorly Designed Data Capture</td>
</tr>
<tr>
<td>Cleansing</td>
<td>Easily Accessible</td>
<td>Data Entry with No Safety Measure/Validation</td>
</tr>
<tr>
<td>Standardization</td>
<td></td>
<td>Cross System Data Flow</td>
</tr>
<tr>
<td>Validation</td>
<td></td>
<td>Coded Patient Data Across Systems</td>
</tr>
<tr>
<td>Enrichment</td>
<td></td>
<td>Security/Accessibility Tradeoff</td>
</tr>
<tr>
<td>Monitoring &amp; Resolution</td>
<td></td>
<td>Lack of Data Lineage/Source Data</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Outcomes from Gaps</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient Bias and Lack of Insight</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poor HCO Financial &amp; Health Outcomes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HCO Interoperability</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
DOWNSTREAM CONSEQUENCES OF BAD DATA
HEALTHCARE STATISTICAL ANALYSIS

I. Research biostatistics
   - Clinical studies
   - Hypothesis testing
   - Comparing “p-values”

II. Healthcare process monitoring statistics
   - Examining rates through time
   - Finding signals vs. normal random variation (“random noise”)
   - Examining signals for root cause
DATA BIAS IN HEALTHCARE
MEDICATION DELIVERY ERRORS

NUMBER OF INCIDENTS

BIAS: DATE OF RECEIPT, NOT DATE OF INCIDENT
DATA BIAS IN HEALTHCARE

Incident Rate (falls) = \frac{\text{Falls}}{\text{Patient Days}} \times 1,000

(Per month)

MONTH

FALL RATE
per 1,000 patient days

Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec

UCL

AVE

CruzStreet
Data Science on Demand
DATA BIAS IN HEALTHCARE

The graph shows a trend over time, labeled as 'SUBGROUP' on the x-axis and 'FALLS' on the y-axis. There is a training event indicated by an arrow. The top line represents the upper control limit (UCL) and the bottom line represents the average (AVE). The graph indicates fluctuations in falls over the specified period.
DATA BIAS IN HEALTHCARE

1. Creating and collecting data
   - Via connected devices
   - Via data entry screens

2. Data handling
   - Where and how it is stored (data at rest)
   - Where and how streaming data is handled (data in motion)
   - Formatting of data to facilitate analysis
   - Data structures; i.e., one source or multiple sources
   - Data aggregation
   - Data access

3. Data analysis
   - Analytics software
   - Descriptive analytics for trending
   - Advanced analytics for prediction; e.g., machine learning
DATA SCIENCE IN PATIENT HEALTHCARE

UNDERSTAND OBTAINING MEDICAL DATA
- MEDICAL RECORDS
- LABORATORY DATA
- TEST RESULTS
- STREAMING OR/ICU

UNDERSTAND DATA HANDLING
- MONITORING
- EVALUATIONS
- PATIENT INTERACTION
- DATA STRUCTURES

PATIENT RECORDS/LABORATORY DATA
CREATE/COLLECT
GOOD DATA

IT/OT
- STORE
- FORMAT
- ACCESS
GOOD DATA

DATA USERS
ANALYZE
GOOD DATA
- TRENDING
- PREDICTING

UNDERSTAND HEALTHCARE DETAILS
- INDIVIDUAL HEALTH
- AGGREGATE HEALTH
- RESEARCH OUTCOMES
- PATIENT EXPECTATIONS
- FINANCIAL PERFORMANCE

* Data integrity
ANALYZE THE DATA BEFORE ITS COLLECTED

1. Rummaging through the data
   - Stratifying & Aggregating it
   - Trending it
   - Business Intelligence software

2. Unsupervised analysis (no predictor variable)
   - Clustering & Pattern recognition
   - Support vector machines
   - Text analysis of written observations
   - Analytical software

3. Supervised analysis (contains a predictor variable)
   - Machine learning
   - Neural networks, Decision trees, Gradient boosting
   - Predictive modeling
   - Advanced analytics software
• Thank you for your time!
• jameseduarte@yahoo.com
• yacine@cruzstreet.com
• jesse@cruzstreet.com
Example Case Study

I. Built Real Time Predictive Model Forecasting Customer Behavior

II. Customer records were of mixed quality due to inconsistent channel data
   ▪ Imagine if channel data were regional such as NYC vs Westchester or more rural county?
   ▪ Inconsistent user interface/data capture (i.e. website) may deliver biased data for region

III. Model outcomes were not good (bias)

IV. To improve model, improved data quality measures through 3rd party service
   ▪ Additionally enriched file with new data to improve data attribution for performance

V. Model was bias because only 30% of customer data/predictive variables were complete

VI. Upon data quality improvement match back on customer file to independent 3rd party data service provider increased by 55% to an 85% match rate

VII. Cleaning, matching, and further enrichment significantly improved model performance and generated highly predictive model for customer behavior
Co-Founders:

Michael Welling, Chair Partner, Meridian Risk Management

Joanne Gere, Executive Director

Westchester Biotech Project

333 Mamaroneck Avenue, #340
White Plains, New York 10605
917-719-2226

www.WestchesterBiotechProject.org

Westchester Biotech Project
Research Community Resources

-a borderless initiative mapping the future for regional and international collaboration

On Twitter we’re @WestchesterBio