Data Collection and Usage: Illegal or Unethical?

Angela Tidwell

SQL Saturday Dallas June 2019
Hi Y’all, I am Angela Tidwell!

@AngelaTidwell
angela@tidwelltidbits.com
www.TidwellTidbits.com

IDERA Ace 2019

Co-Author Stories From The Trenches: Volume 2 from the Let Them Finish series

If you take photos, please share them on twitter #SQLSatDallas
What is legal?

Legalility is based on, concerned with, or permitted by established laws.
What is Ethical?

Ethics is concerned with one of the five values:

1) non-suffering
2) autonomy
3) equality
4) trust
5) character virtue
Serial Killer found at 85 years old

Is it worth lying to land a job?

Is being a whistleblower a good thing?

Is hiding money from your spouse wrong?
Is it legal?

- **Targeting Shoppers**
  - “Target was able to create a national database of tens of thousands of women who were likely pregnant based on their shopping habits.”
  - [https://imaginenext.ingrammicro.com/Trends/September-2017/Four-Retail-Use-Cases-for-Big-Data](https://imaginenext.ingrammicro.com/Trends/September-2017/Four-Retail-Use-Cases-for-Big-Data)

- **Predicting Future Crimes** (recidivism)
  - “Across the nation, judges, probation and parole officers are increasingly using algorithms to assess a criminal defendant’s likelihood of becoming a recidivist”
  - [https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm](https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm)
  - **COMPAS**, an acronym for Correctional Offender Management Profiling for Alternative Sanctions, is a case management and decision support tool developed by Northpointe used by U.S. courts to assess the likelihood of a defendant becoming a recidivist.

- **Using Family Tree DNA to solve crimes**
  - “In certain circumstances, FamilyTreeDNA may be required by law to comply with a valid court order, valid trial, grand jury, subpoena, or search warrant for genetic or personal information.”
  - “If compelled to disclose your Personal Information to law enforcement, we will do our best, unless prohibited by law, to provide you with notice.”
Sneaky ways data is collected

- License Plates
- Cookies
- Heatmaps
- GPS Tracking
- Signal Trackers
- In-Store Wifi
- Credit/Loyalty Cards
- Social Media Activity
- IOT Sensors
- Facial Recognition Cameras
Unethical/Illegal Data Collection

- Facebook & Cambridge Analytica

**Unethical:**

How do you gather a data set of 50+ million people on a closed platform? You take a personality test disguised as a fun conversation starter.

- Create a psychological (and political) ‘profile’ of individuals.
- Save the friend’s list and the urls of friend’s Facebook pages.
- Create a ‘data pool’ around one person.
- Make Assumptions about friends
- Target accordingly

**Illegal:**

Cambridge Analytica took advantage of freely available data, circumvented terms of use and employed questionable mathematical techniques to sell insight which morally shouldn’t have been sold. (political/religious affiliation)
<table>
<thead>
<tr>
<th>By Technology</th>
<th>Beneficial Uses</th>
<th>Questionable Uses</th>
</tr>
</thead>
<tbody>
<tr>
<td>License Plate Readers</td>
<td>Reading passing cars for traffic on highway; police locating stolen car</td>
<td>Used by private detectives; placed on trucks to gather license plate data broadly</td>
</tr>
<tr>
<td>Facial Recognition</td>
<td>Finding potential terrorists at large sporting events</td>
<td>Used by social networking sites to identify members in pictures</td>
</tr>
<tr>
<td>GPS</td>
<td>Location-based coupons; traffic predictions; directions on map</td>
<td>Location-based stalking; iPhone as a homing beacon</td>
</tr>
<tr>
<td>By Context</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Healthcare</td>
<td>Treatment of cancer; health of pregnancy; Google Flu Trends</td>
<td>Discrimination in healthcare and insurance; app knows how fit you are</td>
</tr>
<tr>
<td></td>
<td>Insights into interaction between medications from search terms; insights into hospital spread of infections</td>
<td>Development of a health score from purchase habits and from search terms</td>
</tr>
<tr>
<td></td>
<td>Identifying veterans' potential suicidal thoughts</td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td>Personalizing student instruction</td>
<td>Using data for possible admissions discrimination</td>
</tr>
<tr>
<td></td>
<td>Accountability for performance by school</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Identifying students at risk of dropping out</td>
<td></td>
</tr>
<tr>
<td>Electricity</td>
<td>Turning on/off home electricity</td>
<td>Allowing criminals to know if you are home; smart homes hacked</td>
</tr>
<tr>
<td>Law Enforcement</td>
<td>Machine learning to identify burglar; accessing phone records to identify potential suspects in a mugging</td>
<td>Accessing smartphone without a warrant; identifying suspects by web browsing habits</td>
</tr>
<tr>
<td></td>
<td>New York Fire Department using data mining to predict problems</td>
<td>Individuals under scrutiny for not participating in tracking</td>
</tr>
<tr>
<td>Retail</td>
<td>Improving layout of store based on typical movements of customers</td>
<td>Tracking movements/shopping habits of spectators at a stadium using Verizon's Precision Marketing Insight program</td>
</tr>
<tr>
<td></td>
<td>Better coupons, suggested items</td>
<td>Price discrimination (e.g., Amazon, Orbitz)</td>
</tr>
<tr>
<td></td>
<td>WalMart's use of RetailLink to integrate suppliers with onsite supplier inventory</td>
<td>Target sending notice of pregnancy to unsuspecting teen's parents</td>
</tr>
<tr>
<td>Urban Planning</td>
<td>Traffic management; smart grid technology</td>
<td>Identifying who is listening to which radio station; EZ Pass responder tracked everywhere</td>
</tr>
<tr>
<td></td>
<td>Use of popular app by competitive cyclists and runners for road planning</td>
<td>Possibility of hackers changing traffic lights and creating traffic jams</td>
</tr>
<tr>
<td></td>
<td>Identifying areas for road improvement</td>
<td>Identifying areas for road improvement but focusing only on those with mobile apps</td>
</tr>
</tbody>
</table>

@AngelaTidwell
### Beneficial/Questionable Uses

#### By Technology

<table>
<thead>
<tr>
<th>By Technology</th>
<th>Beneficial Uses</th>
<th>Questionable Uses</th>
</tr>
</thead>
<tbody>
<tr>
<td>License Plate Readers</td>
<td>Reading passing cars for tolls on highway; police locating stolen car</td>
<td>Used by private detectives; placed on trucks to gather license plate data broadly</td>
</tr>
<tr>
<td>Facial Recognition</td>
<td>Finding potential terrorists at large sporting events</td>
<td>Used by social networking sites to identify members in pictures</td>
</tr>
<tr>
<td>GPS</td>
<td>Location-based coupons; traffic predictions; directions on map</td>
<td>Location-based stalking; iPhone as a homing beacon</td>
</tr>
</tbody>
</table>
## Beneficial/Questionable Uses

### By Context

<table>
<thead>
<tr>
<th>By Context</th>
<th>Beneficial Uses</th>
<th>Questionable Uses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthcare</td>
<td>Treatment of cancer; health of pregnancy; Google Flu Trends; Insights into interaction between medications from search terms; insights into hospital spread of infections; Identifying veterans’ potential suicidal thoughts</td>
<td>Discrimination in healthcare and insurance; app knows how fit you are; Development of a health score from purchase habits and from search terms</td>
</tr>
<tr>
<td>Education</td>
<td>Personalizing student instruction; Accountability for performance by school; Identifying students at risk of dropping out</td>
<td>Using data for possible admissions discrimination</td>
</tr>
<tr>
<td>Electricity</td>
<td>Turning on/off home electricity</td>
<td>Allowing criminals to know if you are home; smart homes hacked</td>
</tr>
</tbody>
</table>
## Beneficial/Questionable Uses By Context (cont.)

<table>
<thead>
<tr>
<th>Category</th>
<th>Beneficial Uses</th>
<th>Questionable Uses</th>
</tr>
</thead>
</table>
| **Law Enforcement** | Machine learning to identify burglars; accessing phone records to identify potential suspects in a mugging  
New York Fire Department using data mining to predict problems | Accessing smartphone without a warrant; identifying suspects by web browsing habits  
Individuals under scrutiny for not participating in tracking |
| **Retail**      | Improving layout of store based on typical movements of customers  
Better coupons, suggested items  
Walmart’s use of RetailLink to integrate suppliers with onsite supplier inventory | Tracking movements/shopping habits of spectators at a stadium using Verizon’s Precision Marketing Insight program  
Price discrimination (e.g., Amazon, Orbitz)  
Target sending notice of pregnancy to unsuspecting teen’s parents |
| **Urban Planning** | Traffic management; smart grid technology  
Use of popular app by competitive cyclists and runners for road planning  
Identifying areas for road improvement | Identifying who is listening to which radio station; EZ Pass responder tracked everywhere  
Possibility of hackers changing traffic lights and creating traffic jams  
Identifying areas for road improvement but focusing only on those with mobile adds |
Things Are Not Always as They Seem
Bias in Data

- Human Bias
  - Conscious Bias (aka Explicit Bias)
  - Unconscious Bias (aka Implicit Bias)
- Machine Bias
  - Sample Bias
  - Prejudice Bias
  - Measurement Bias
  - Algorithm Bias
Hidden Prejudices

- Name
- Client ID
- DOB
- Zip Code
- Race
- Gender
- Sexual Orientation
- Education Level
- Religion
- Spending levels
- Items bought at stores
Biased Data Can Skew Results

**Sample Bias:**
HS students surveyed to measure use of drugs (does not include homeschooled, private schooled, or drop-outs)

*Cherry picking for “best” results.

**Prejudice Bias:**
ABC reported a new ABC News-Washington Post poll that showed President Trump’s approval rating “ticking up slightly” to 40%. The on-screen graphic also pointed out that 56% disapprove. (40 is not 56, but it makes for a better story)

**Measurement Bias:**
Poll asks Atlanta Falcon’s fans questions about Drew Brees’ abilities as a quarterback. Results will be skewed away from science and more toward their feelings.

**Algorithm Bias:**
A team of all white males create an algorithm to find the perfect candidate for a position as “Director of Diversity”
Reducing Bias

Ways to reduce bias in data:

- Use a diverse group to collect and code the data.
- Diverse group testing and review.
- Verify results with more data sources.
- Be critical, check for alternative explanations.
- Identify outliers.
- Identify source of collection.
- Confirm findings with peers.
References

- EdX.ORG Microsoft MPP Course DAT249X Ethics & Law in Data & Analytics
  https://courses.edx.org/courses/course-v1:Microsoft+DAT249x+2T2018/course/
  https://medium.com/@melodyucros/10-sneaky-ways-companies-are-collecting-data-to-understand-customers-beob9089d54a
  https://privacyinternational.org/campaigns/investigating-apps-interactions-facebook-android
  https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm
  https://www.familytreedna.com/legal/terms-of-service

©AngelaTidwell
Thank you!

Thank you so much for your time and attention today!

Please feel free to reach out to me:

@AngelaTidwell

angela@tidwelltidbits.com

TidwellTidbits.com