Disinformation Defense of AI Inference

KAREL BALOUN, KEN CHANG, MATT HOLMES

MASTER OF INFORMATION AND CYBERSECURITY (MICS)

UNIVERSITY OF CALIFORNIA – BERKELEY

DECEMBER 12, 2019

Freedom On The Line

Freedom is America's seminal and central value.

"A threat to free will is a threat to freedom, the imposition of a dangerous worldview without public awareness. **When free will itself is threatened, that is the ultimate threat to freedom**."

George Lakoff, UC Berkeley

on page 62 of "Whose Freedom?" (2006)

Private Voter Data Risk

- Personalized messaging damages election integrity
- Campaign regulation and privacy law are behind
- Our ML model prototypes behavioral prediction
- AI speed and optimization, utilizing personal information, will soon pose new dramatic threats
- 5 Recommendations: PoDD-BAm

AI or Fly

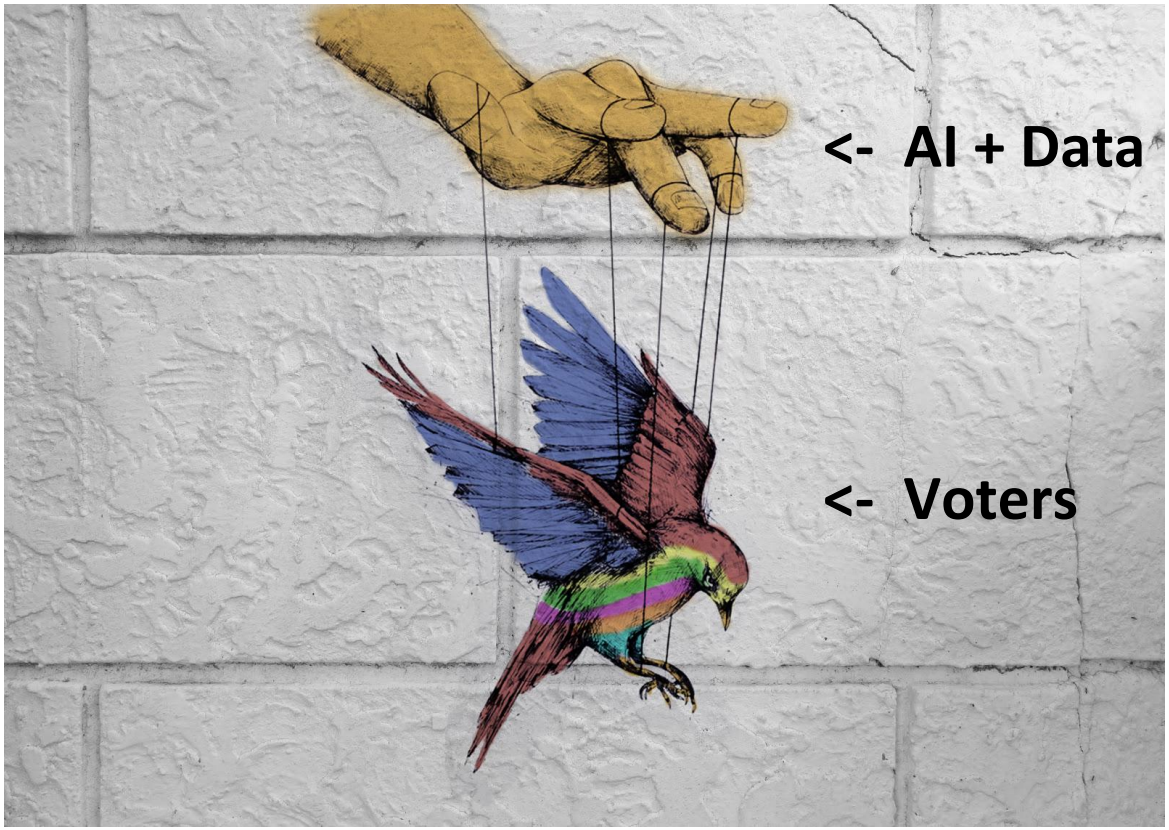


image credit: https://c.wallhere.com/photos/ee/fe/birds_colorful_puppets_freedom-1229625.jpg!d

Threat Modeling

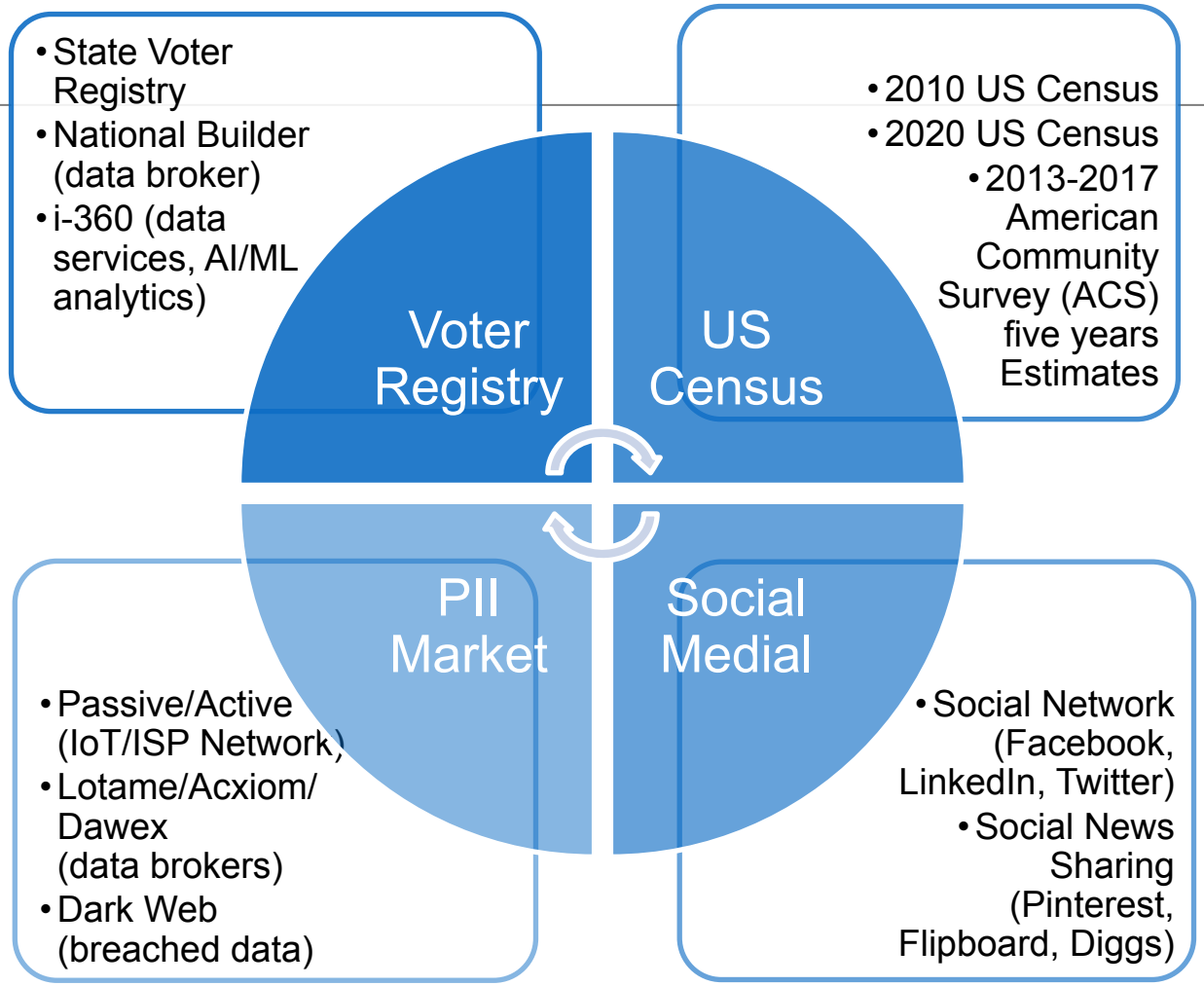
Minimal Voter Data: Identify & Protect

Identify Voters
Legal Name
Age
Residential Address
Party

Protect Fair Elections
Voted, or not
*Ethnicity**
*Gender**
*Polling Location**

**These attributes are candidates for privacy engineering!*

Pillars of Personal Data



Inference Threats

"Concerns about algorithmic accountability are often actually concerns about the way in which **these technologies draw privacy invasive and non-verifiable inferences about us that we cannot predict, understand, or refute.**“

Sandra Wachter and Brent Mittelstadt

of the Oxford Internet Institute at University of Oxford

Threat Modeling

Assets	Threats	Description
Voter Database	Confidentiality, Integrity	Public static database contains personal attributes such as full legal name, birth date, gender, residential address and registered party affiliation. Voter data broker provides services for distribution with fee and might subject to data poisoning attack.
US Census Database	Confidentiality	Public database contains summary statistics of personal attributes such as ethnicity, income, education, marriage, and number of children. Legally required to protect personal information. With combination of synthetic data generation, data aggregation and linkage, may leak sensitive personal attributes like vote preference.
Social Media	Confidentiality	Data leakage via public post with personal attributes, personal association with degrees of separation, and subject to targeted marketing advertisement.
Personal Data Market	Confidentiality, Integrity	Collective Data breach events creates black market of personal data in Dark Web. Consumer service provider collects and resell the personal attributes for profit or in exchange of free service. Personal data exchange market promotes monetized model for individual. The personal data broker might subject to data poisoning attack.
Personal Privacy	Integrity	Personal service right might subject to perpetuate discrimination because machine learning result are trained on biased data. An individual might not receive the service they needed.

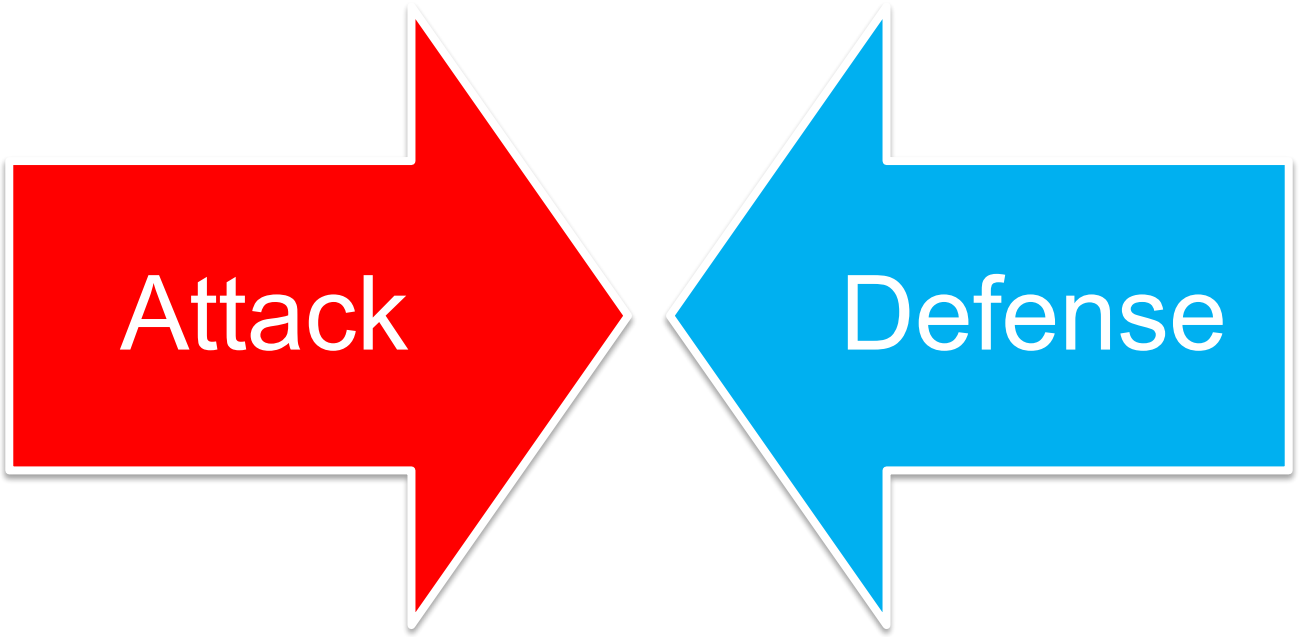
Messaging & AI Threats

- Messaging Bots & Email Spam
- Robot calls, including for push surveys
- Media blogs, videos and posts
- Advertising by PACs, Issue, and Campaigns

- AI impersonation, lie creation.
- Overloading the media and messaging, especially locally
- Not big data challenge, rather persuasion management + intensity
- “Her”. Speed and interactivity: of natural language responses
- Tuning of persuasive messaging (Cambridge Analytica at max scale)

Our Approach

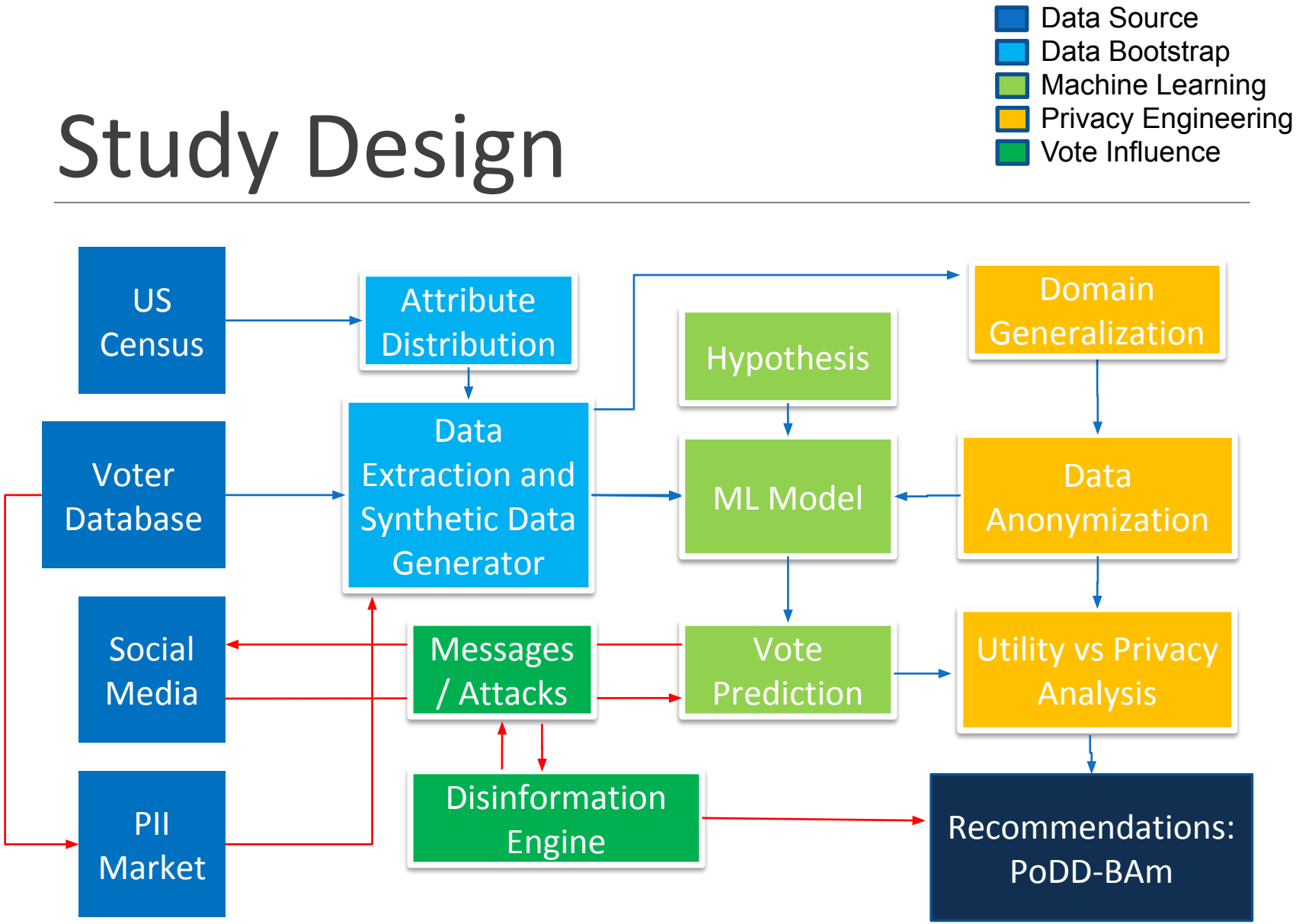
Hypothesis:
Voter Preferences Disclosure



Use of Machine Learning algorithm with vast among of personal data to infer secret personal attribute on vote preference.

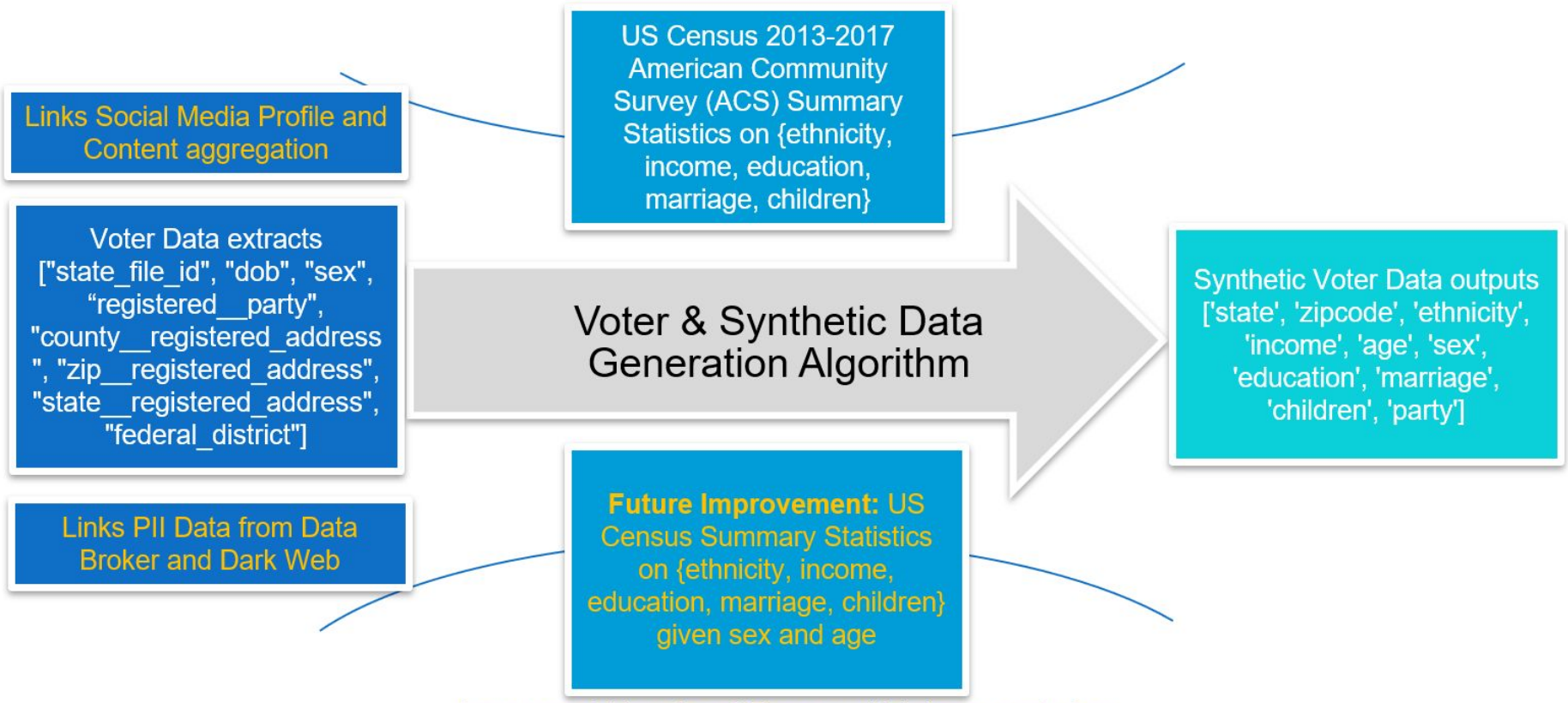
Advocate on personal data privacy laws and regulations with holistic approach on anonymize the personal data in the public market.

Study Design

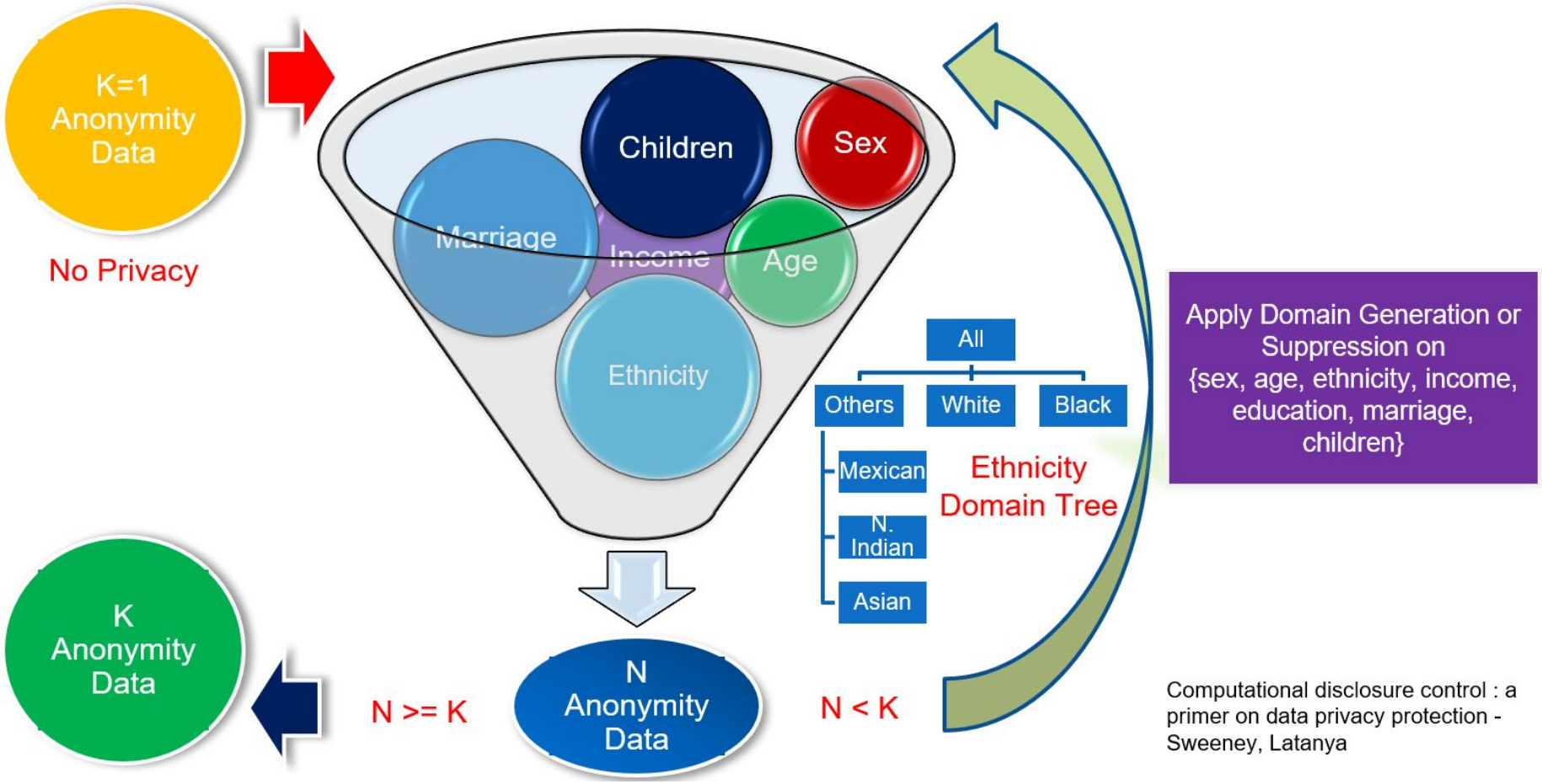


Defense: Privacy Engineering

Voter Data Bootstrapping



Voter Data Anonymization -
Datafly Algorithm with desired K-Anonymity

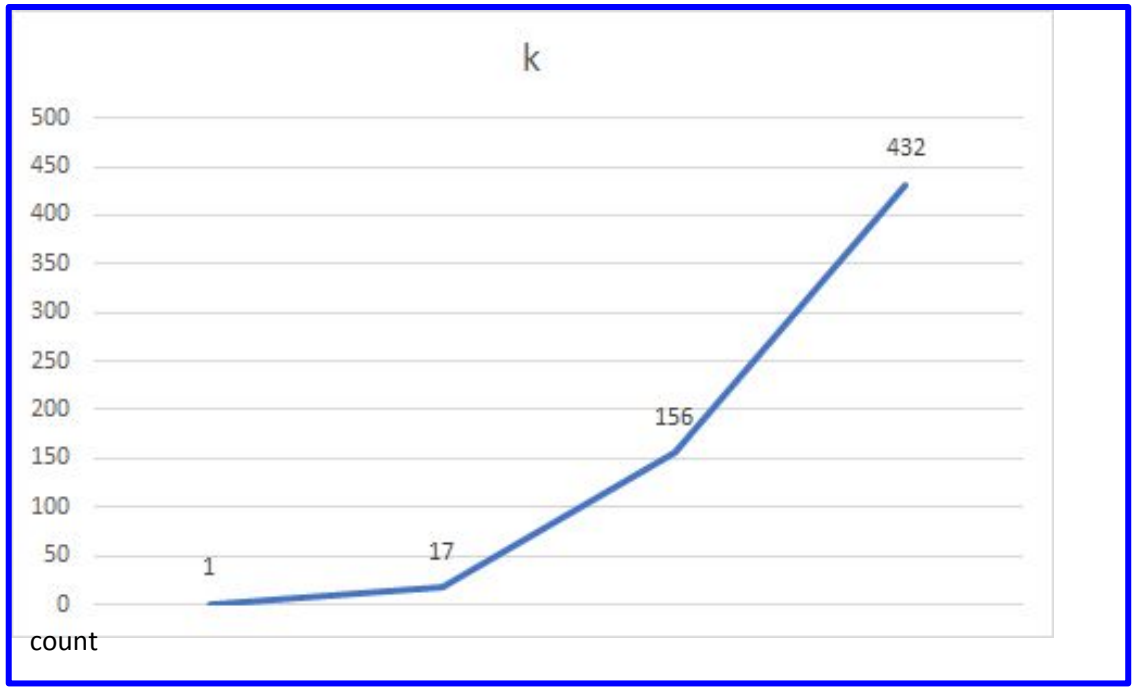


Computational disclosure control : a primer on data privacy protection - Sweeney, Latanya

Voter Data Privacy vs. Utility

qID = {'ethnicity', 'income', 'age', 'sex', 'education', 'marriage', 'children'};

k=1 on the raw data mixed from voter data and synthetic attributes

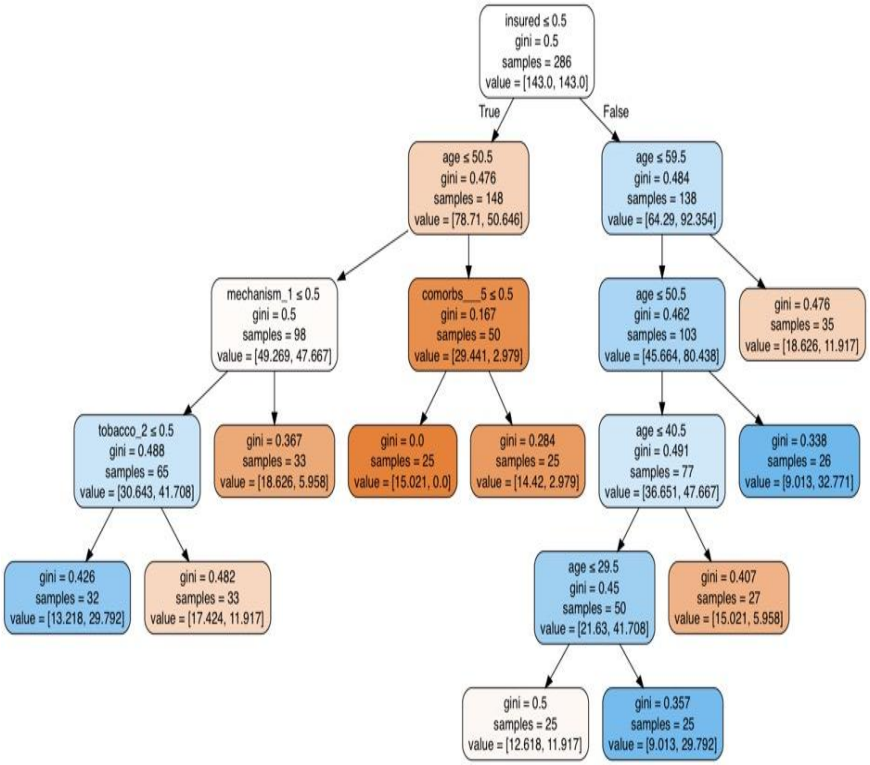


Attack: Machine Learning

ML Model - Decision Tree

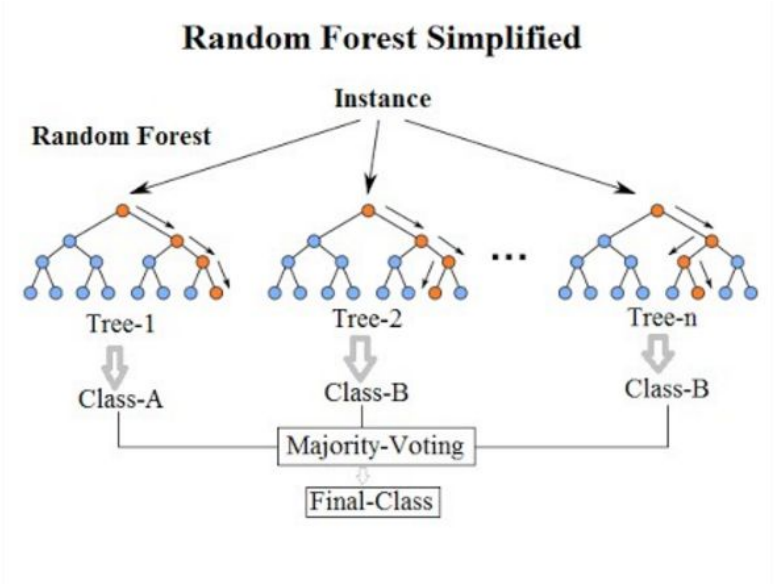
Supervised Model - Party as Labels (as a proxy for vote preference)

- Decision Tree
- Train & Predict raw data
  - `Raw Data: 40% acc`
- Train & Predict anonymized data
  - `Anonymized Data: 47% acc`
- Baseline: 3 labels ~ 33%
- Summarize Results



ML Model - Privacy Aware Decision Forest

Decision Forest

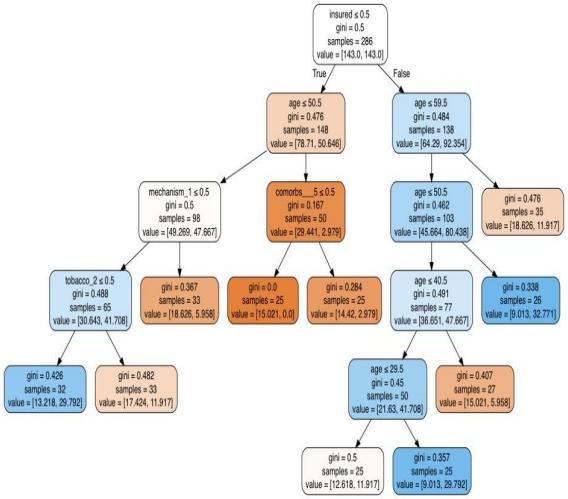


Privacy Aware

- Each decision tree will be tuned with Privacy Engineering in mind
- Increased accuracy when combined with more carefully constructed synthetic data
- More robust than a single tree; will help with capturing more structure in the data

Summary of “Privacy Aware”

Ex: Max Leaf Count



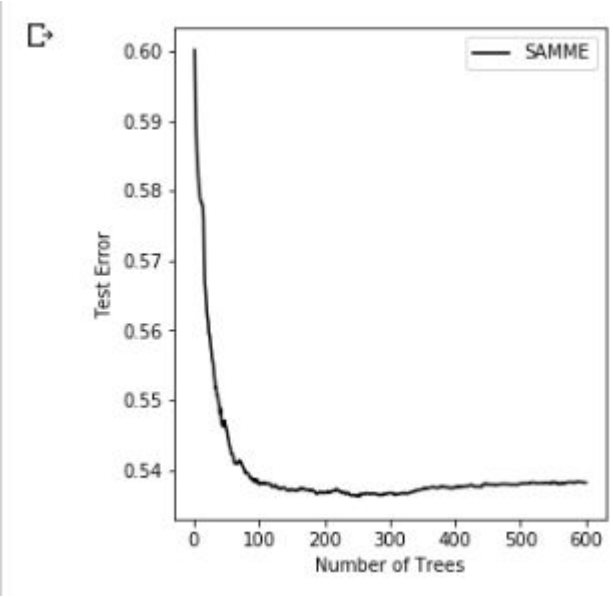
- Leaf Count ≤ Number of ECs

Ex: Carefully synthesized data

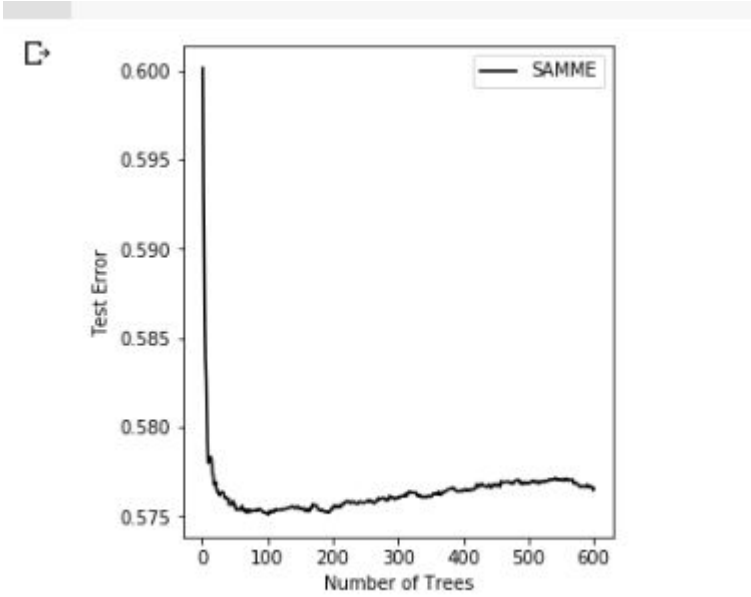
1. Use Census to collect probabilities of attributes for each EC
2. Find datasets that tie attributes to party affiliation (even in population level)

Forest Results

K=1 Data: 46% accuracy



K=17 Data: 43% accuracy



Limitations and Resolutions

Limitation

1. Full data
2. Using synthetic attributes for the sake of the model
3. Decision Tree/Forest may be too weak to learn the patterns

Resolution

1. Purchase, or otherwise compile more complete data
2. Be more careful in constructing synthetic data to capture more predictive value
3. Try additional classifiers; eg, KNN, SVM, Neural Net

Lessons Learned

PoDD-BAm: 5 Recommendations

- Further study of **Privacy by Design** principles **on** collection, processing, and disclosure of personal voter and public census data.
- Regulate the **data broker** industry with consumer privacy law, so everyone can view and opt-out any sensitive information.
- Every Secretary of State should mandate **disclosure requirements** for access and use of **voter record files**.

- FEC must **ban use voter personal profiles** in voter messaging, and ban use completely by campaigns and PACs.
- Large **audience minimum** size for advertising.

Next Steps/Roadmap

Buy/Build private profile data, to show precise benefits of hiding it via privacy engineering.

Demonstrate ML generated messages, as cheap, belief-conforming, attention-attracting voter spam.

Iterate to improve the **ML voter sentiment and behavior prediction model**, to measure the impact of information attacks.

RSA Conference 2020 - Topic Accepted; Draft Submission on Jan 7th

image credit: Wordclouds.com


Q
A

Credits

A Big Thank You! to:

- **Dr. Ebrima Ceesay** & **Dr. Nathan Good**, for this class!
- **Lisa Ho**, for introductions, motivation, organization
- **Daniel Aranki**, for PE and ML guidance.
- **Matt Bishop** and **Carlos Rivera**, for the data landscape
- **Hany Farid**, for fake video and fake news guidance
- **Anonymous Source** formerly at Cambridge Analytica
- Our **MICS classmates**, and this Program…
  we've all been here for each other.

Supporting Material

More Next Steps

- Continue to flesh out write-up for final format
- Try to identify source of inaccuracy in ML model
  - Overfitting
  - Anonymization
- Perform multiple cycles of anonymization & model training
- Optimize anonymization
- Approach additional conferences

Personal Data : definition

According to the law, personal data means any information relating to an identified or identifiable individual; an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number (e.g. social security number) or one or more factors specific to his physical, physiological, mental, economic, cultural or social identity (e.g. name and first name, date of birth, biometrics data, fingerprints, DNA...)

Commission Nationale de l'Informatique et des Libertés. cnil.fr

Inference Threats

"These inferences draw on highly diverse and feature-rich data of unpredictable value, and create new opportunities for discriminatory, biased, and invasive decision-making. Concerns about algorithmic accountability are often actually concerns about the way in which these technologies draw privacy invasive and non-verifiable inferences about us that we cannot predict, understand, or refute.“

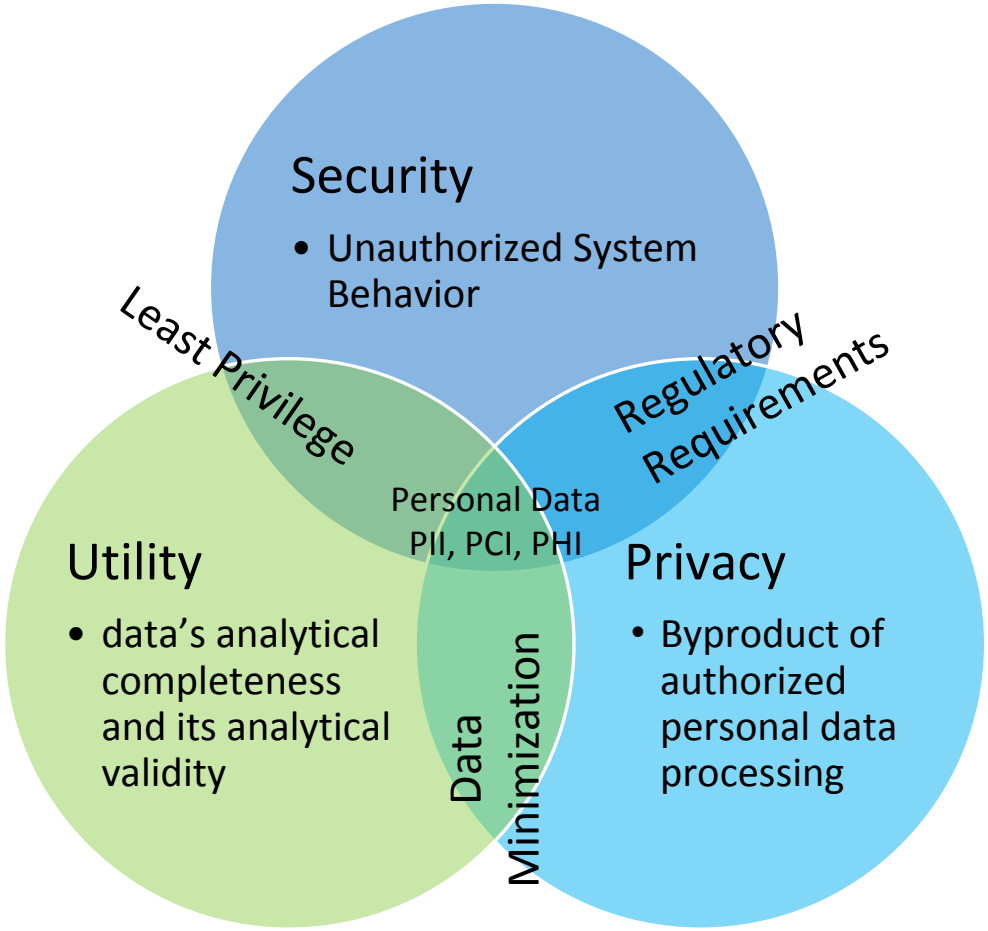
Sandra Wachter and Brent Mittelstadt

of the Oxford Internet Institute at University of Oxford

Disinformation Engine HOWTO

- "Disinformation" may be impossible to define and ban directly.
- Strongest attacks **reinforce voters existing beliefs**, or persuade voters to **do nothing**.
- Probabilistic profile matching was the greatest achievement of Cambridge Analytica. (not psychometric analysis, nor info theft)
- Political ads are only one vector. Coordinated posts, all media and personal messaging are all even more dangerous.
- AI can optimize messaging: the **big data is all possible messaging**.
- AI can **iterate and improve**: organize, maximize message volume.

Security vs. Privacy vs. Utility



Tug of War

sportycious.com

US Census Attribute Distribution Algorithm

- 2018 estimates available from 2010 census data, for income, education, family size, and ethnicity.
- These can be filtered by geographic area, gender, and age.
- Data is difficult to extract from web interface. API exists, but getting an apikey includes some delay, and sample code limited.
- Margin of error can rapidly overwhelm any information content, as filters lower the number of class members.
- Due to margin error, not possible to map to voter database, beyond large sample proportional distributions.

- Real world individual voter profiles exist, with all of these attributes and many more. These exist as purchasable data, as well as advertising API targets.

Real Election Case Studies

- Brexit: easy propagation of false messages, confirmation bias
- Hong Kong: false stories of anti-police violence, state media
- Canada: bots and networks reused
- Trump: Russia Guccifer hack, GRU media, IRA social

- Zuckerberg's dilemma
- US freedom of speech, vs UK libel law, etc.
- China & Russia's complete control of media messages

Domain Generalization

qID = {'ethnicity', 'income', 'age', 'sex', 'education', 'marriage', 'children'};

- ethnicity
  - {0,1,2,3,4,5}
  - {0} -> {20} -> {30}
  - {1} -> {21} -> {30}
  - {3,4,5} -> {22} -> {30}
- income
  - {0,1,2,3,4,5,6}
  - {0} -> {10}-> {20} -> {30}
  - {1,2} -> {11}-> {20} -> {30}
  - {3} -> {12}-> {20} -> {30}
  - {4} -> {13}-> {21} -> {30}
  - {5,6} -> {14}-> {21} -> {30}
- sex
  - {0,1,2}
  - {0,1} -> {10}-> {20}
  - {2} -> {11}-> {20}
- education
  - {0,1,2,3,4,5,6,7}
  - {0}-> {20}-> {30}-> {40}
  - {1,2,3}-> {21}-> {30}-> {40}
  - {4,5}-> {22}-> {31}-> {40}
  - {6,7}-> {23}-> {31}-> {40}
- marriage
  - {0,1,2,3}
  - {0} -> {20} -> {30}
  - {1} -> {21} -> {30}
  - {3,4} -> {22} -> {30}
- children
  - {0,1,2,3,4,5}
  - {0} -> {20} -> {30}
  - {1,2} -> {21} -> {30}
  - {3,4,5} -> {22} -> {30}

Conclusions

image credit: Dante's Inferno & https://www.flickr.com/photos/spine/2432632767

Bounds on Quality

- Discernibility metric (CDM) assigns to each tuple t in V a penalty, which is determined by the size of the equivalence class containing t

$$C_{DM} = \sum_{EquivClasses\ E} |E|^2$$

- The normalized average equivalence class size metric (C_{AVG}) measures how well partitioning approaches the best case.
- This metric means that the quality of anonymized data is measured by the average size of equivalence classes

$$C_{AVG} = \left(\frac{total\_records}{total\_equiv\_classes} \right) / (k)$$

- C_{AVG} is to reduce the normalized average equivalence class size*

*nih.gov

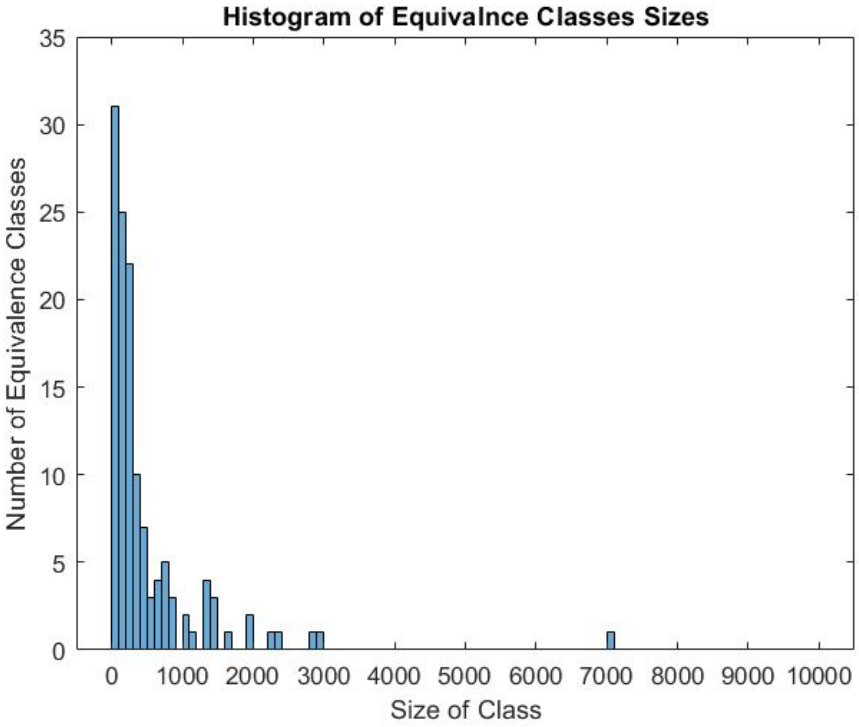
Synthetic Attribute Distribution (US Census)

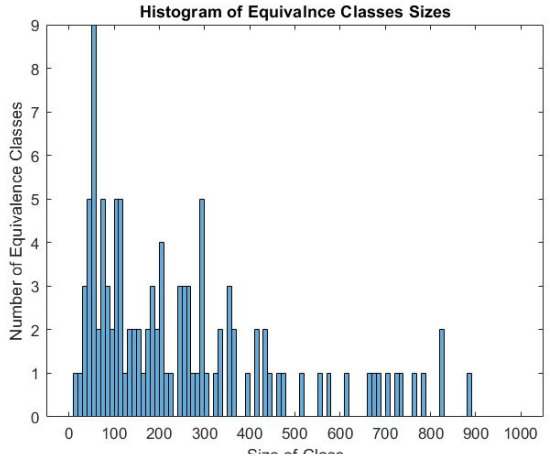
```
# IA District 1 personal data distribution per US Census (by Karel)
a_ethnicity = ['NA','White','Black','Mexican','Native Indian','Asian']
p_ethnicity = [0.0,0.92,0.05,0.005,0.005,0.02]

a_income = ['0-14999', '15000-34999', '35000-49999', '50000-74900', '75000-99999',
'100000-149000', '150000-']
p_income = [0.11,0.2,0.13,0.2,0.14,0.12,0.1]

a_education =
['NA','elementary','middle','high','associate','bachelor','master','phd']
p_education = [0,0.03,0.05,0.33,0.32,0.19,0.07,0.01]

a_marriage = ['widowed','not married', 'married', 'divorced']
p_marriage = [0.08,0.3,0.5,0.12]

a_children = ['NA', 'one', 'two', 'three', 'four', 'five']
p_children = [0.7,0.2,0.07,0.01,0.01,0.01]
```

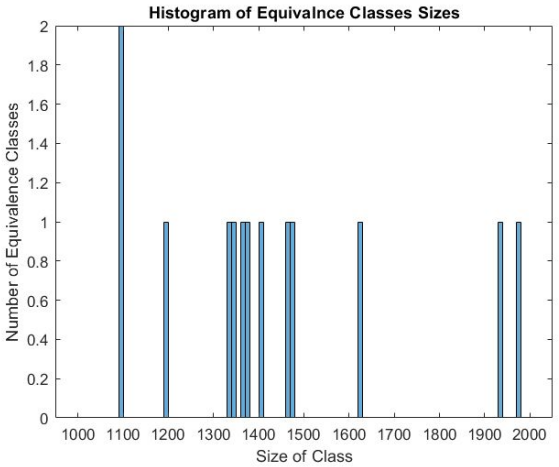
Disclosure Attacks

- Identity disclosure: being able to tell the identity of the person to whom the record corresponds. For example, telemonitoring project collected data on zip code and birthday with de-identified username. The linking (join) of telemonitoring data and voter data leads to identity disclosure.
- Attribute disclosure: being able to tell that a person has a specific (sensitive) attribute. For example, based on the average estimated distance and general description of terrain, the adversary could infer the health status (sensitive attribute).
- Membership disclosure: type of threat based on multiple queries and background Information. For example, identify a data subject is part of telemonitoring project.

Data Anonymization (K=17)







Data Bootstrapping

Input: Voter Data File, vd_file_name, Voter File Table, VFT: vf_attributes = ["state_file_id", "dob", "sex", "party", "county__registered_address", "zip__registered_address", "state__registered_address", "federal_district"], Synthetic Attribute Distribution, p_[ethnicity, income, education, marriage, children]

Output: Data Output Table: dt_attributes = ['state', 'zipcode', 'ethnicity', 'income', 'age', 'sex', 'education', 'marriage', 'children', 'party']

Algorithm:

```
vf_data = load_VF_data(vd_file_name, vf_attributes)

sd_ethnicity = generateEthnicity(len(vf_data), p_ethnicity)
sd_income = generateIncome(len(vf_data), p_income)
sd_education = generateEducation(len(vf_data), p_education)
sd_marriage = generateMarriage(len(vf_data), p_marriage)
sd_children = generateChildren(len(vf_data), p_children)

 for i, row_ in enumerate(vf_output):
        for index in range(0, n):
                dt_row = generate_columns(index, row_, dt_attributes)
        append_row(dt_row)

output(dt_output_file)
```

Future Enhancement:

Given Age & Sex attributes from voter database, the quality of synthetic data attributes {ethnicity, income, education, marriage, children} can be improved with summary statistics group by age and sex.

Datafly Algorithm

Core Datafly Algorithm

Input: Private Table PT; quasi-identifier QI = $(A_1, \ldots, A_n)$, k-anonymity constraint k; domain generalization hierarchies DGH_{A_i}, where i=1,…,n with accompanying functions f_{A_i}, and *loss*, which is a limit on the percentage of tuples that can be suppressed. PT[*id*] is the set of unique identifiers (key) for each tuple.

Output: MGT a generalization of PT[QI] that enforces k-anonymity

Assumes: $|PT| \geq k$, and *loss* * $|PT| = k$

algorithm Datafly:

// Construct a frequency list containing unique sequences of values across the quasi-identifier in PT,
// along with the number of occurrences of each sequence.

1. **let** freq be an expandable and collapsible Vector with no elements initially.Each element is of the form (QI, *frequency*, *SID*), where $SID = \{id_i : \exists t[id] \in PT[id] \Rightarrow t[id] = id_i\}$; and, *frequency* = |*SID*|. Therefore, freq is also accessible as a table over (QI, *frequency*, *SID*).
2. **let** *pos* ← 0, *total* ← 0
3. **while** total ≠ |PT| **do**
   - 5.1 freq[*pos*] ← (*t*[QI], *occurs*, *SID*)
     where *t*[QI]∈PT[QI], (*t*[QI],__, ___)∉ freq; *occurs* = |PT| - |PT[QI] – {*t*[QI]}|;
     and, $SID = \{id_i : \exists t[id] \in PT[id] \Rightarrow t[id] = id_i\}$
   - 5.2 *pos* ← *pos* + 1, *total* ← *total* + *occurs*

// Make a solution by generalizing the attribute with the most number of distinct values
// and suppressing no more than the allowed number of tuples.

6. **let** *belowk* ← 0
7. **for** *pos* ← 1 **to** |freq| **do**
   - 7.1 (__, *count*) ← freq[*pos*]
   - 7.2 **if** *count* < *k* **then do**
     - 7.2.1 *belowk* ← *belowk* + *count*
8. **if** *belowk* > k **then do:** *// Note. loss* * |PT| = k
   - 8.1 freq ← **generalize**(freq)
   - 8.2 **go to** step 4
9. **else do**

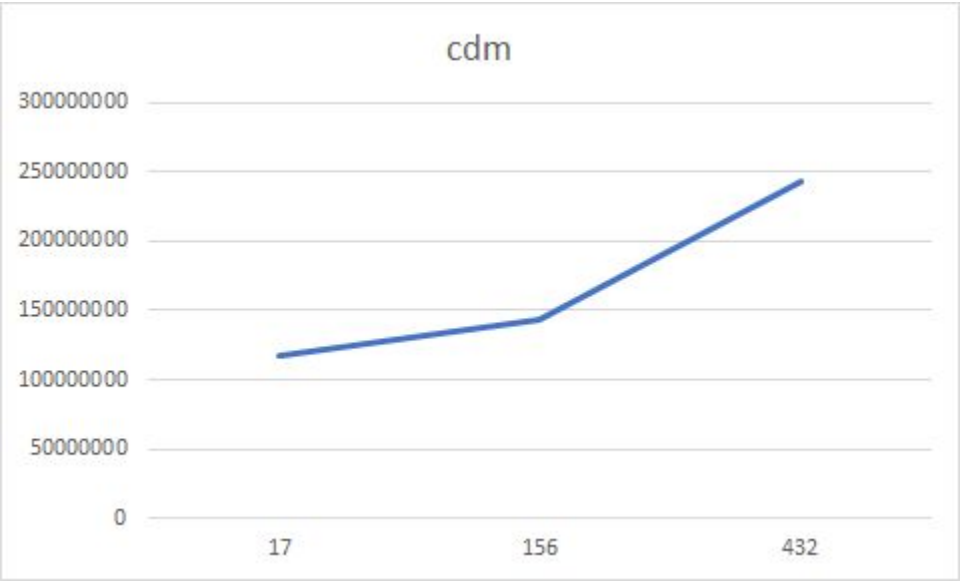
// assert: the number of tuples to suppress in freq *is* ≤loss * |PT|

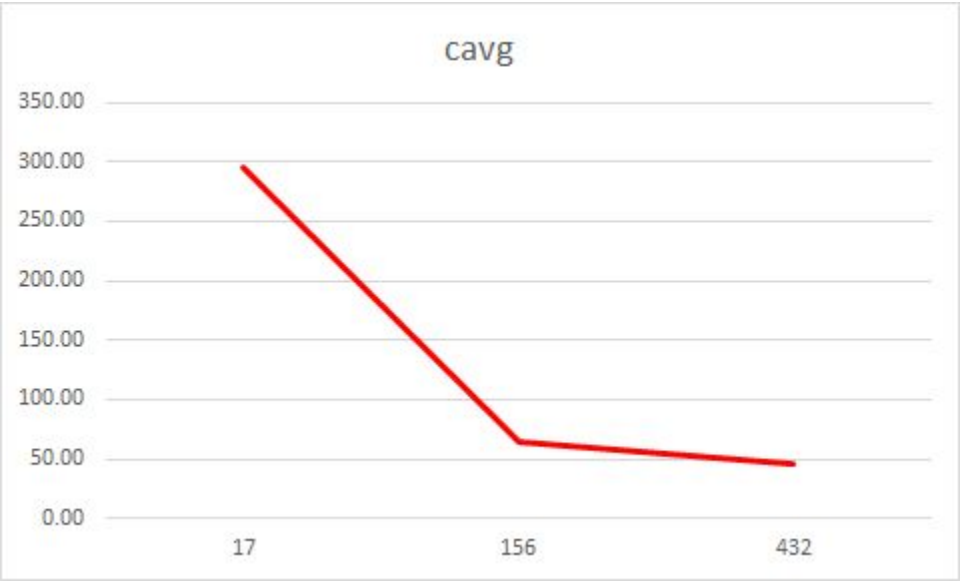
   - 9.1 freq ← **suppress**(freq, *belowk*)
   - 9.2 MGT ← **reconstruct**(freq)
10. **return** MGT.

```
~$ python3 datafly.py -pt "IA_synthetic_USCensusDist_full_data_Output.csv" -qi
"ethnicity" "income" "age" "sex" "education" "marriage" "children" -dgh
"ethnicity_generalization.csv" "income_generalization.csv" "age1_generalization.csv"
"sex_generalization.csv" "education_generalization.csv" "marriage_generalization.csv"
"children_generalization.csv" -k 30 -o "k_30_anon_IA_USCensusDist_data_full.csv"
[LOG] Created output file.
...
[LOG] Current attribute with most distinct values is 'age'.
[LOG] Generalizing attribute 'age' for sequence 9111...
[LOG] Generalized attribute 'age'. Current generalization level is 3.
[LOG] 17530 tuples are not yet k-anonymous...
...
[LOG] Current attribute with most distinct values is 'income'.
[LOG] Generalizing attribute 'income' for sequence 4830...
[LOG] Generalized attribute 'income'. Current generalization level is 2.
[LOG] 12662 tuples are not yet k-anonymous...
...
[LOG] Current attribute with most distinct values is 'education'.
[LOG] Generalizing attribute 'education' for sequence 1781...
[LOG] Generalized attribute 'education'. Current generalization level is 2.
[LOG] 7464 tuples are not yet k-anonymous...
...
[LOG] Current attribute with most distinct values is 'children'.
[LOG] Generalizing attribute 'children' for sequence 838...
[LOG] Generalized attribute 'children'. Current generalization level is 2.
[LOG] 2157 tuples are not yet k-anonymous...
...
[LOG] Current attribute with most distinct values is 'ethnicity'.
[LOG] Generalizing attribute 'ethnicity' for sequence 127...
[LOG] Generalized attribute 'ethnicity'. Current generalization level is 2.
[LOG] 0 tuples are not yet k-anonymous...
[LOG] Suppressed 0 tuples.
[LOG] Writing anonymized table...
```

Computational disclosure control : a primer on data privacy protection - Sweeney, Latanya

Discernibility Cost vs. Average Equivalence Class Size



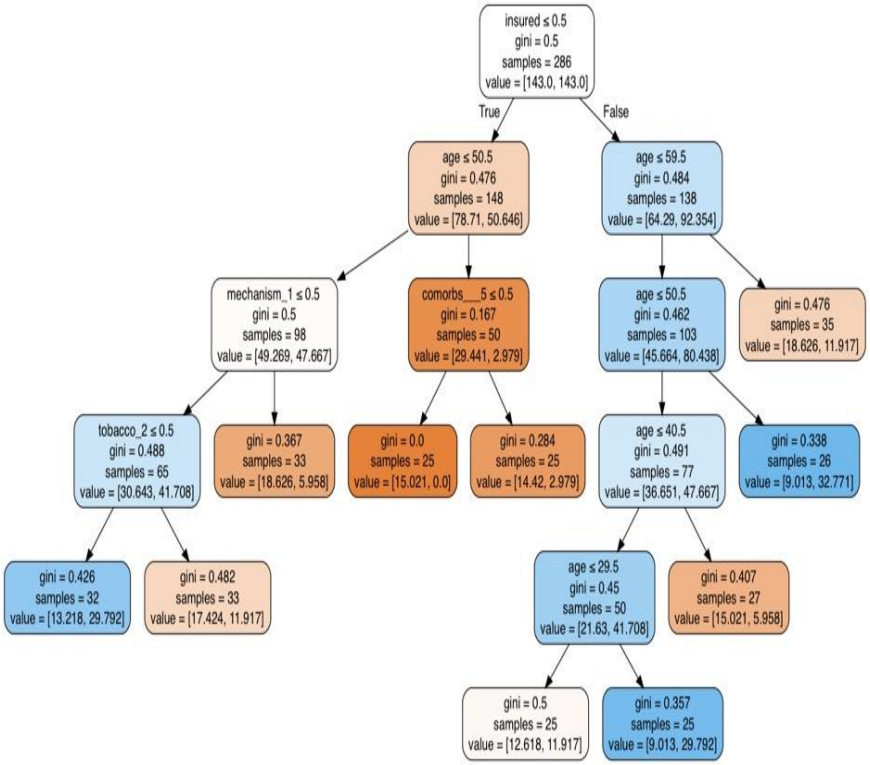


ML Extras

ML Model - Decision Tree

Supervised Model - Synthetic Labels

- Decision Tree
- Train & Predict raw data
  - `Real data: 65% acc`
- Train & Predict anonymized data
  - `Anonymized data: 78% acc`



More on Privacy Aware - max leaf count

- For Decision Trees in general, the decision to predict a label based on your attributes depends on what leaf your record falls into
- From privacy engineering, the records with the same attributes form an equivalence class
- So it follows that records within the same equivalence class would be classified into the same leaf

- So it doesn't make sense to have more leaves than equivalence classes

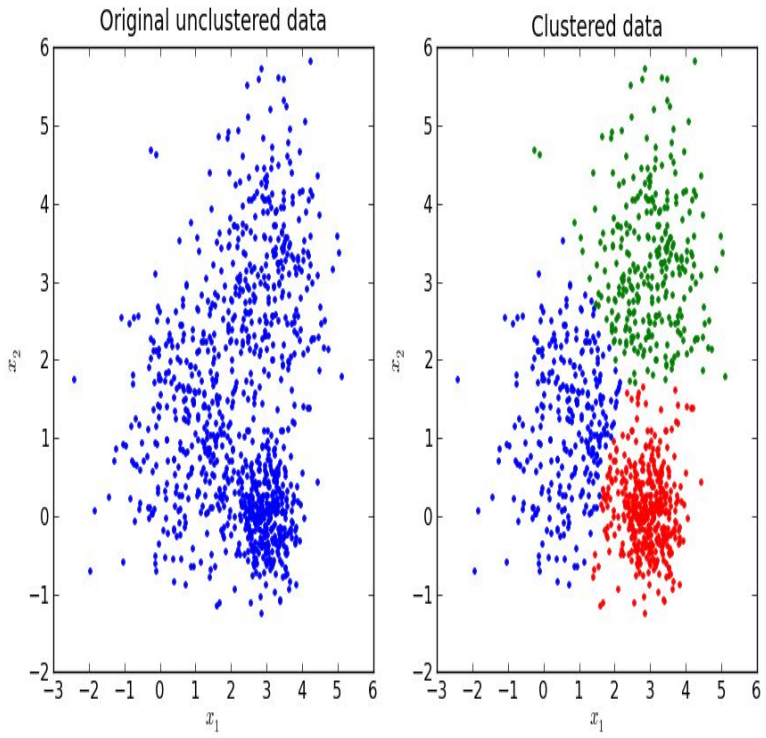
More on Privacy Aware - carefully synthesized data

- Currently, our synthesized attributes are based on population level probabilities
  - For example, P(MS=Married) = (Number Married / Total Number)
  - This assumes that all attributes are completely independent of one another; which they aren't
- More carefully constructed probabilities would be done per Equivalence class of non-synthetic attributes
  - For example, P(MS = Married | Age, Gender)
  - Increases accuracy by capturing more structure to the data
- Even more carefully constructed probabilities would include the label
  - For example, P(MS = Married | Age, Gender, Party)
  - This leverages the fact that the attribute is actually tied to the label even if it is still only at the population level

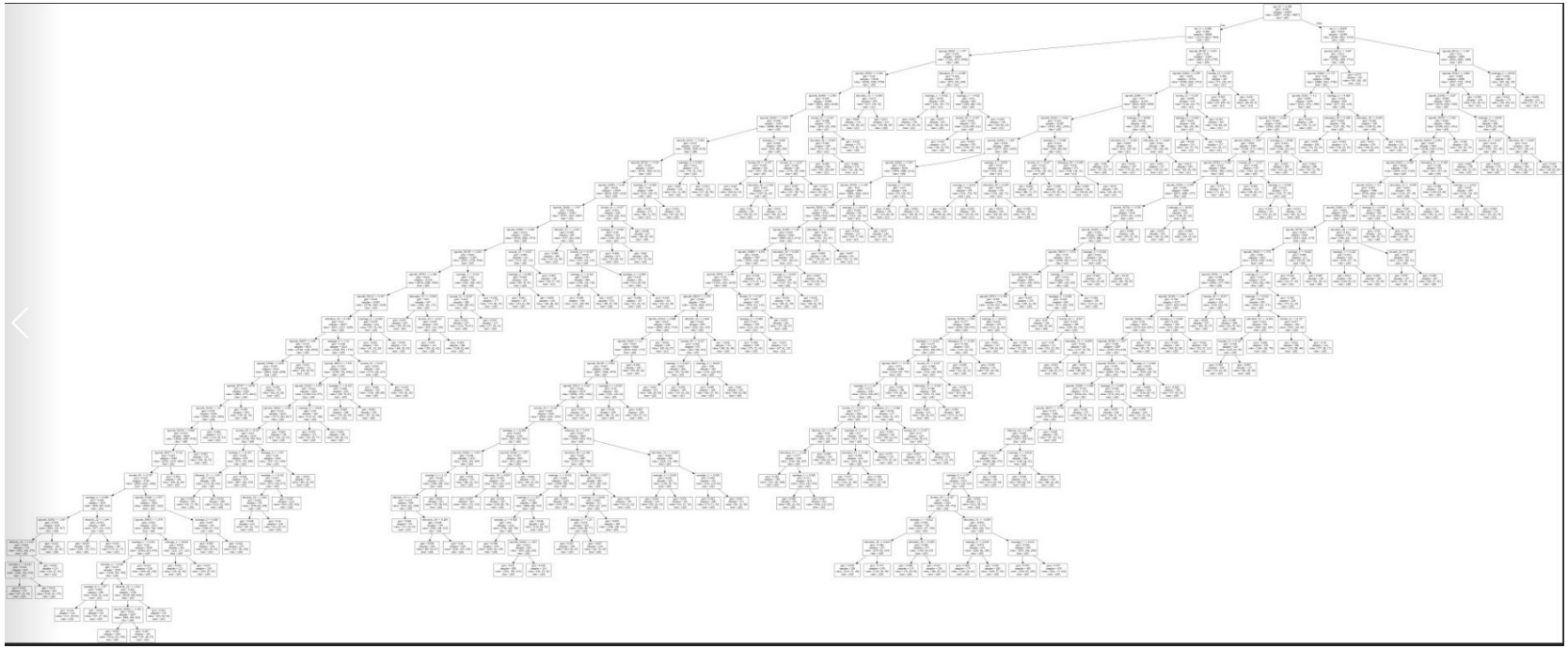
ML Model - Kmeans

Party included

- Decision Tree
- Train & Predict raw data
  - `None - 0.2701415799204721`
  - `Dem - 0.6626522807284148`
  - `Rep - 0.06720613935111314`
- Train & Predict anonymized data
  - `None - 0.2701415799204721`
  - `Dem - 0.6626522807284148`
  - `Rep - 0.06720613935111314`



Anonymized Tree



Raw Tree