What is Privacy?

- “The right to be let alone” - Warren & Brandeis, 4 Harvard L.R. 193 (Dec. 15, 1890)
  - My information protected so it doesn’t adversely affect me in the future
- Control over data
  - My information used only in ways I approve
- Issues:
  - Disclosure / sharing
  - Approved use
  - Recourse
Data Privacy: The Goal

- Protect the Individual
  - “Everyone has the right to the protection of personal data concerning him or her. Such data must be processed fairly for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law. Everyone has the right of access to data which has been collected concerning him or her, and the right to have it rectified.” – Charter of Fundamental Rights of the European Union

- Challenges: What do we mean by
  - “concerning” an individual
  - Protection
  - Consent
  - Access / rectified

“Obvious” answers

- Concerning an individual
  - Has your name/address/other identifying information

- Protection
  - Only used/accessed in expected, intended, authorized ways

- Consent
  - You know and agree to what is done with the data

- Access/Rectify
  - You can see the data and correct errors
Consent?

The Guardian
Maev Kennedy
Thu 11 Jun 2009 07.17 EDT

American family’s web photo ends up as Czech advertisement

Smiths from Missouri only heard about it when a friend travelling in Prague saw them on a grocery store poster

Could facebook have done this?

Facebook didn’t authorize it, it but could they?

Facebook Terms of Service 4/19/18: when you share, post, or upload content that is covered by intellectual property rights (like photos or videos) (exclusive, transferable, sublicensable, royalty-free, and worldwide license to host, use, distribute, modify, run, copy, publicly perform or display, translate, and create derivative works of your content (consistent with your privacy and application settings). This means, for example, that if you share a photo on Facebook, you give us permission to store, copy, and share it with others (again, consistent with your settings) such as service providers that support our service or other Facebook Products you use.

Before 4/19/18, if shared with others, deleting your account didn’t terminate these rights.
“Obvious” answers

• Concerning an individual
  – Has your name/address/other identifying information

• Protection
  – Only used/accessed in expected, intended, authorized ways

• Consent
  – You know and agree to what is done with the data

• Access/Rectify
  – You can see the data and correct errors

Concerning an Individual:

IC 24-4.9-2-10 (Breach Disclosure)

IC 24-4.9-2-10 "Personal information"

Sec. 10. "Personal information" means:
(1) a Social Security number that is not encrypted or redacted; or
(2) an individual's first and last names, or first initial and last name, and one (1) or more of the following data elements that are not encrypted or redacted:
   (A) A driver's license number.
   (B) A state identification card number.
   (C) A credit card number.
   (D) A financial account number or debit card number in combination with a security code, password, or access code that would permit access to the person's account.

The term does not include information that is lawfully obtained from publicly available information or from federal, state, or local government records lawfully made available to the general public.

Other codes (e.g., spyware prohibition) have different definitions
The AOL Awakening

• In Aug 2006, AOL released its customers web searches for research studies
• 20 Million unique queries of 650K unique users
• <user-id> was replaced with a <random number>
• NY Times reporter successfully found the identity of an individual from the queries
  – Queries included “60 single men” “landscapers in Lilburn, Ga”
  – Many more queries contained enough information to uniquely identify the person
• And it keeps going (Netflix, NYC Taxi, …)

AOL fired its CTO over this issue;
Two researchers were forced out

Re-identifying “anonymous” data
(Sweeney ’01)

• 37 US states mandate collection of information
• Dr. Sweeney purchased the voter registration list for Cambridge Massachusetts
  – 54,805 people
• 69% unique on postal code and birth date
• 87% US-wide with all three

Solution: k-anonymity
  – Any combination of values appears at least k times
• Developed systems that guarantee k-anonymity
  – Minimize distortion of results
Anonymity: The Goal

- Prevent Disclosure of Personal Information
  - GDPR: ‘personal data’ means any information relating to an identified or identifiable natural person (‘data subject’); an identifiable natural person is one who can be identified, directly or indirectly
  - Qatar Law 13 of 2016: Personal Data: Data belonging to an Individual with specified or reasonably specifiable identity whether through such Personal Data or through combining the same with any other data
  - *But still use the data where appropriate!*
- Problem: It can’t be done!
  - “Perfect” privacy requires zero utility (e.g., the data must be encrypted.)
  - As soon as we can use the data (e.g., decrypt), it is at risk

Why Perfect Privacy is Impossible

*(Dwork, McSherry, Nissim, and Smith ’06)*

- Background Knowledge
  - Adversary may already know a lot
  - Whatever we provide (even de-identified or anonymized data) may add to that knowledge
- It may just take that “last bit of knowledge” to give the adversary the ability to violate privacy
  - *We can formally prove 1 bit may be too much*
- The possibility is real
What We Can Do

- Encryption
  - Reduce risk to minimal levels when data not in use
- Anonymization
  - Produce usable data that is hard to link to individuals
- Noise addition
  - Usable data where any link to individuals (or information we surmise about individuals) is guaranteed to be uncertain/suspect

What We Can Do: Encryption

- Goal: Reduce risk to minimal levels when data not in use
- Encrypted Computation
  - Process the data while it is encrypted
  - Decrypt final output: Generalized, non-individual results
- Basic tools
  - Homomorphic Encryption, Commutative Encryption, Order Preserving Encryption
- Research Prototypes can accomplish many data processing and analysis tasks using these tools
  - Garbled Computing: Compute without revealing either the data or the program

Garbled Computing.

Software pub. or client
Source code & data
Optimizing Compiler
Garbled code
Garbled output
Virtual GC

GC runs as a black box on one or more host computers (some are untrusted) in a cooperative & non-colluding fashion
What We Can Do: Anonymization

- Ensure protected/sensitive data not directly identifiable
  - Remove links between protected data and identifiers
- Generalize “quasi-identifiers”: Information that when combined with external data enables re-identification
  - Birth dates, addresses, workplace, etc.
  - E.g., instead of birth date, only give year
- Anonymized data still useful for data analysis
  - Goal is general knowledge, not learning specifics about individuals
- Example: “Anatomized” database from “Private Data in the Cloud” project

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>ID</th>
<th>Manufacturer</th>
<th>Drug Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roan</td>
<td>1</td>
<td>Raphe Healthcare</td>
<td>Retinoic Acid</td>
</tr>
<tr>
<td>Lisa</td>
<td>2</td>
<td>Raphe Healthcare</td>
<td>Retinoic Acid</td>
</tr>
<tr>
<td>Roan</td>
<td>3</td>
<td>Envie De Neuf</td>
<td>Mild Exfoliation</td>
</tr>
<tr>
<td>Elyse</td>
<td>4</td>
<td>Emedoutlet</td>
<td>Nexium</td>
</tr>
<tr>
<td>Carl</td>
<td>5</td>
<td>Jai Radhe</td>
<td>Adapalene</td>
</tr>
<tr>
<td>Roan</td>
<td>6</td>
<td>Hangzhou Btech</td>
<td>Cytarabine</td>
</tr>
<tr>
<td>Lisa</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Roan</td>
<td>8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

What We Can Do: Noise Addition

- Idea: Impact of noise on what we learn from the data larger than impact of any individual's data
- Formally: For $S \subseteq Range(f)$, an $\epsilon$-differentially private mechanism $M$ satisfies
  $$\frac{Pr[M_f(D_1) \in S]}{Pr[M_f(D_2) \in S]} \leq e^\epsilon$$
  where $D_1$ and $D_2$ differ on at most one element
- U.S. Census Bureau is starting to use Differential Privacy

$D$

$f(D) = 17$

$M_f(D_1)$

$M_f(D_2)$

$e^\epsilon$
Achieving Differential Privacy

• Laplace Mechanism
  – Add Laplacian noise to the query result
  – Calibrate noise to the sensitivity of the query
    \[ \text{Private } f(D) = f(D) + \text{Lap}\left(\frac{\Delta f}{\epsilon}\right) \]

• Sensitivity
  – Captures the largest contribution to the result that can be made by one individual
    \[ \Delta f = \max_{D,D'} |f(D) - f(D')| \]

Another Example: Randomized Response (Warner ’65)

• For each respondent with a yes/no value \( f(D) \), flip a coin:
  – If heads, \( \text{Private } f(D) = f(D) \)
  – If tails, \( \text{Private } f(D) = \text{a second coin flip} \)

• True answer
  \[ \text{avg } f(D) = \frac{1}{4} (1 - \text{avg Private } f(D)) + \frac{3}{4} \text{avg Private } f(D) \]

• Differentially private with \( \epsilon = \ln 3 \)
  – Changing first coin flip changes epsilon
Exponential Mechanism

- The exponential mechanism $M_E(x, u, R)$ selects and outputs an element $r \in R$ with probability proportional to
  \[
  \exp\left(\frac{\epsilon u(x, r)}{2\Delta u}\right)
  \]
  - $x$ is database, $u$ captures how much a given $r$ distorts the outcome for the database $x$
  - $\Delta u$ is sensitivity – maximum distortion $r$ can cause across neighboring databases

Cool Properties of Differential Privacy

- Assume $M$ is a differentially private mechanism
- Post-processing: $f \circ M$ is differentially private
  - Once the results are differentially private, anything we do with the results (that doesn’t look back at the data) is still private
- Composition: $M_1, M_2$ are $\epsilon_1, \epsilon_2$-differentially private mechanisms
  - $M_1(x), M_2(x)$ is $(\epsilon_1 + \epsilon_2)$-differentially private
  - If $x, y$ disjoint, $M_1(x), M_2(y)$ is $\max(\epsilon_1, \epsilon_2)$-differentially private

*Some caveats on this*
privacy parameter

• How to set $\epsilon$
  – Open problem
  – No rule of thumb or guidelines
  – Physical meaning of $\epsilon$
  – indistinguishable = unidentifiable?

(a) large $\epsilon$  (b) small $\epsilon$

Differential Identifiability
*(Lee&Clifton, KDD’12)*

• Issue: What is the right value for $\epsilon$?
  – Tells how far the answer is off
  – Want to bound probability of identification:
    \[ \Pr[i \in D \mid M_f(D)=R] \leq \rho \]

• Differential Privacy easy enough to achieve
  – Adding Laplacian noise guarantees $\epsilon$-differential privacy
    \[ M_f(X) = f(X) + \text{Lap}\left(\frac{\Delta f}{\epsilon}\right) \]
  – Somewhat more complicated for Differential Identifiability
    • But same basic approach/math
**Identifiability**

- Privacy game
  - $\mathcal{U} = \{u_1, u_2, \ldots, u_m\}$

  1. Pick a database $D \in \mathcal{U}^n$
     - $D = (d_1, \ldots, d_n)$
     - $D' = D - \{d_n\}$
  2. Query $f$
  3. $\mathcal{M}(D) = f(D) + \text{Lap}\left(\frac{\Delta f}{\epsilon}\right)$
  4. Send $(D', r = \mathcal{M}(D))$

- To limit adversary’s confidence to $\rho$, what value of $\epsilon$ should we use?

**Differential identifiability**

- Practical definition
  - privacy based on differential privacy
  - probabilistic interpretation of individual identifiability

- Definition
  - $\mathcal{M}$ satisfies $\rho$-differential identifiability if
    $$\forall D' = D - \{i\}, \forall i \in U - D'$$
    $$\Pr[I(i) \in I_D | \mathcal{M}_f(D) = R, D'] \leq \rho$$
privacy parameter

- Simple example
  - U={1, 2, 3, 4, 5, 10}
  - D = {1, 2, 3, 10} , D’={1, 2, 3}
  - f=mean, R=5.4

<table>
<thead>
<tr>
<th>ψ</th>
<th>f</th>
<th>ϵ = 1</th>
<th>ϵ = 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>{1,2,3,4}</td>
<td>2.5</td>
<td>0.2353</td>
<td>0.1478</td>
</tr>
<tr>
<td>{1,2,3,5}</td>
<td>2.75</td>
<td>0.2666</td>
<td>0.1898</td>
</tr>
<tr>
<td>{1,2,3,10}</td>
<td>4</td>
<td>0.4981</td>
<td>0.6624</td>
</tr>
</tbody>
</table>

Table 1. Pr[D = ω_i | R, D’]

- DP-classifier [Cormode KDD 2011]
  - build an ε-DP naïve bayes classifier
  - can predict (potentially) sensitive information w.h.p.

Differential identifiability

- Assumption
  - uniform prior distribution
  - or, \( t = \max_i \Pr[\psi_i = D] \) is known

- Relationship to DP
  - Any \( \epsilon \)-differential private mechanism satisfies
    \[
    \frac{1}{1 + (m-1)e^{-\epsilon}} \text{-differential identifiability}
    \]
  - inherits nice properties of DP
  - graceful degradation
revisiting the example

- Setting
  - $\rho = 0.4$
  - $\lambda = \frac{s(f)}{\ln(\frac{1}{1-\rho})} = 5.21$

- Possible worlds
  - $\omega_1 = \{1, 2, 3, 4\}$
  - $\omega_2 = \{1, 2, 3, 5\}$
  - $\omega_3 = \{1, 2, 3, 10\}$
  - $\Pr[M_f(\omega_1) = R] = 0.0589$
  - $\Pr[M_f(\omega_2) = R] = 0.0618$
  - $\Pr[M_f(\omega_3) = R] = 0.0785$
  - $\Pr[\omega_3 = D | R] = \frac{0.0785}{0.0589 + 0.0618 + 0.0785} = 0.3941$

---

**What We Can Do: Anonymization**

- Ensure protected/sensitive data not directly identifiable
  - Remove links between protected data and identifiers

- Generalize “quasi-identifiers”: Information that when combined with external data enables re-identification
  - Birth dates, addresses, workplace, etc.
  - E.g., instead of birth date, only give year

- Anonymized data still useful for data analysis
  - Goal is general knowledge, not learning specifics about individuals

- Example: “Anatomized” database from “Private Data in the Cloud” project

<table>
<thead>
<tr>
<th>ID</th>
<th>Manufacturer</th>
<th>Drug Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Raphe Healthcare</td>
<td>Retinoic Acid</td>
</tr>
<tr>
<td>2</td>
<td>Raphe Healthcare</td>
<td>Retinoic Acid</td>
</tr>
<tr>
<td>3</td>
<td>Raphe Healthcare</td>
<td>Retinoic Acid</td>
</tr>
<tr>
<td>4</td>
<td>Envie De Neuf</td>
<td>Mild Exfoliation</td>
</tr>
<tr>
<td>5</td>
<td>Emuoutlet</td>
<td>Nexium</td>
</tr>
<tr>
<td>6</td>
<td>Gep-Tek</td>
<td>Abiraterone</td>
</tr>
<tr>
<td>7</td>
<td>Jai Radhe</td>
<td>Adapalene</td>
</tr>
<tr>
<td>8</td>
<td>Hangzhou Btech</td>
<td>Cytarabine</td>
</tr>
</tbody>
</table>

---
Problems with Anonymity

- Can we really prevent re-identification?
  - Experience says no
  - Big Data (*Variety*) makes it worse
- If we can, is the data still useful?
  - Currently having serious issues with anonymizing city-sized health information dataset

Myth: Anonymity is Broken

- Many academic papers with attacks on anonymization
  - E.g., deFinetti (*Kifer’09*), Minimality (*Wong, Fu, Wang, Pei ’07*)
  - Real-world failures (e.g., AOL)
- Reality: There is a risk
  - But risk may be acceptable (e.g., HIPAA safe-harbor rules do not eliminate risk of re-identification)
  - Differential Privacy provides provable limits on risk
  - **Any disclosure that provides utility also carries some privacy risk** (*Dwork’06*)
**$\ell$-Diversity**

- Example using Bucketization
  - Anatomy (Xiao et al. (2006))

<table>
<thead>
<tr>
<th>Age (A)</th>
<th>Zipcode (Z)</th>
<th>Job (J)</th>
<th>GID (G)</th>
</tr>
</thead>
<tbody>
<tr>
<td>41</td>
<td>47905</td>
<td>Assoc. Prof</td>
<td>1</td>
</tr>
<tr>
<td>29</td>
<td>47906</td>
<td>Assist. Prof</td>
<td>1</td>
</tr>
<tr>
<td>30</td>
<td>47906</td>
<td>Assist. Prof</td>
<td>2</td>
</tr>
<tr>
<td>35</td>
<td>47907</td>
<td>Assoc. Prof</td>
<td>2</td>
</tr>
<tr>
<td>28</td>
<td>47906</td>
<td>Assist. Prof</td>
<td>3</td>
</tr>
<tr>
<td>47</td>
<td>47905</td>
<td>Prof.</td>
<td>3</td>
</tr>
<tr>
<td>45</td>
<td>47905</td>
<td>Prof.</td>
<td>4</td>
</tr>
<tr>
<td>31</td>
<td>47906</td>
<td>Assist. Prof</td>
<td>4</td>
</tr>
</tbody>
</table>

Identifier Table (IT)

<table>
<thead>
<tr>
<th>GID (G)</th>
<th>Income (I)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>[100K-150K]</td>
</tr>
<tr>
<td>1</td>
<td>[50K-75K]</td>
</tr>
<tr>
<td>2</td>
<td>[75K-100K]</td>
</tr>
<tr>
<td>2</td>
<td>[50K-75K]</td>
</tr>
<tr>
<td>3</td>
<td>[75K-100K]</td>
</tr>
<tr>
<td>3</td>
<td>[100K-150K]</td>
</tr>
<tr>
<td>4</td>
<td>[100K-150K]</td>
</tr>
<tr>
<td>4</td>
<td>[75K-100K]</td>
</tr>
</tbody>
</table>

Sensitive Table (ST)

- $k$-Anonymous and $\ell$-Diverse
- Every instance in IT can be matched to $\ell=2$ instances in ST.

**HIPAA: De-Identifying Data**

- A person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable
  - Applying such principles and methods, determines that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information; and
  - Documents the methods and results of the analysis that justify such determination
- The following identifiers of the individual or of relatives, employers, or household members of the individual, are removed:
  - Names, Location < 1st three digits of zip, dates < year,
    Tel/Fax/email/SSN/MRN/InsuranceID/Account/licence/VIN/License Plate Numbers, DeviceID, URL/IP, Biometric IDs, full-face photographs, any other unique identifiers; and
  - The covered entity does not have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a subject of the information.
Anonymized Data

- HIPAA Safe-Harbor De-Identified Data
  - Is it useful?

<table>
<thead>
<tr>
<th>Name</th>
<th>Addr.</th>
<th>Birth</th>
<th>Sex</th>
<th>Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>479xx</td>
<td>56</td>
<td>F</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>479xx</td>
<td>67</td>
<td>M</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>479xx</td>
<td>78</td>
<td>M</td>
<td>Schizophrenic</td>
<td></td>
</tr>
</tbody>
</table>
Anonymized Data

• HIPAA Safe-Harbor De-Identified Data
  – Is it useful?
  – Is it enough?

<table>
<thead>
<tr>
<th>Name</th>
<th>Addr.</th>
<th>Birth</th>
<th>Sex</th>
<th>Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>479xx</td>
<td>56</td>
<td>F</td>
<td>…</td>
<td></td>
</tr>
<tr>
<td>479xx</td>
<td>67</td>
<td>M</td>
<td>…</td>
<td></td>
</tr>
<tr>
<td>479xx</td>
<td>78</td>
<td>M</td>
<td>Schizophrenic</td>
<td></td>
</tr>
</tbody>
</table>
Anonymized Data

- HIPAA Safe-Harbor De-Identified Data
  - Is it useful?
  - Is it enough?

<table>
<thead>
<tr>
<th>Name</th>
<th>Addr.</th>
<th>Birth</th>
<th>Sex</th>
<th>Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>479xx</td>
<td>56</td>
<td>F</td>
<td>Uses Marijuana for Phantom Pain</td>
<td></td>
</tr>
<tr>
<td>479xx</td>
<td>67</td>
<td>M</td>
<td>Uses Marijuana for Pain</td>
<td></td>
</tr>
<tr>
<td>479xx</td>
<td>78</td>
<td>M</td>
<td>Schizophrenic</td>
<td></td>
</tr>
</tbody>
</table>

Myth: Anonymized Data Isn’t Useful

- Differential Privacy: Noise added for privacy is often small relative to other sources of noise in the data
  - Can even improve statistical value of results (Dwork et al. ‘17)
- Machine Learning models learned from Anonymized Data can be as good or better than using the original data
  - Decision trees on k-anonymous data (Iyengar’02)
  - Support Vector Machines on anonymized data (Mancuhan&Clifton’17)
  - Nearest Neighbor on anonymized data
Machine Learning from Anonymized Data (Mancuhan & Clifton’17)

- Binary Classification task: predict an attribute in IT given the other attributes in IT and the attribute in ST
  - Example: predict age <35 or >=35 given job, zipcode and income
- What about predicting the attribute in ST table? (Example: income)
  - Amounts to defeating privacy
- Why do we care about using ST?
  - Income may be useful to predict Job

<table>
<thead>
<tr>
<th>Age (A)</th>
<th>Zipcode (Z)</th>
<th>Job (J)</th>
<th>GID (G)</th>
</tr>
</thead>
<tbody>
<tr>
<td>41</td>
<td>47905</td>
<td>Assoc. Prof</td>
<td>1</td>
</tr>
<tr>
<td>29</td>
<td>47906</td>
<td>Assist. Prof</td>
<td>1</td>
</tr>
<tr>
<td>30</td>
<td>47906</td>
<td>Assist. Prof</td>
<td>2</td>
</tr>
<tr>
<td>35</td>
<td>47907</td>
<td>Assoc. Prof</td>
<td>2</td>
</tr>
<tr>
<td>28</td>
<td>47906</td>
<td>Assist. Prof</td>
<td>3</td>
</tr>
<tr>
<td>47</td>
<td>47905</td>
<td>Prof.</td>
<td>3</td>
</tr>
<tr>
<td>45</td>
<td>47905</td>
<td>Prof.</td>
<td>4</td>
</tr>
<tr>
<td>31</td>
<td>47906</td>
<td>Assist. Prof</td>
<td>4</td>
</tr>
</tbody>
</table>

Identifier Table (IT)

<table>
<thead>
<tr>
<th>GID (G)</th>
<th>Income (I)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>[100K-150K)</td>
</tr>
<tr>
<td>1</td>
<td>[50K-75K)</td>
</tr>
<tr>
<td>2</td>
<td>[75K-100K)</td>
</tr>
<tr>
<td>2</td>
<td>[50K-75K)</td>
</tr>
<tr>
<td>3</td>
<td>[75K-100K)</td>
</tr>
<tr>
<td>3</td>
<td>[100K-150K)</td>
</tr>
<tr>
<td>4</td>
<td>[100K-150K)</td>
</tr>
<tr>
<td>4</td>
<td>[75K-100K)</td>
</tr>
</tbody>
</table>

Sensitive Table (ST)

Learning from Anonymized Data

- Anatomization: Possible to learn accurate models to classify data
- Can even outperform the models that are trained on original data in terms of
  - Error Rate (K-NN, Linear SV Classifier)
  - Convergence (1-NN)
- Can also reduce error compared to using attributes in IT alone
- Much better and practical than models for generalized/suppressed data
- Large training set helps…
**Use Cases for Anonymity**

- **Public release**
  - Challenging, given possible attacks on anonymity
- **Protection from “accidental re-identification”**
  - Ethical researchers could see their neighbor...
  - Model: HIPAA Limited Dataset
    - Easily re-identifiable, but only released under Data Use Agreement
- **Reduce risk from data breach**
  - Which would you rather have stolen, identifiable data or anonymized, possibly re-identifiable data
  - *Won’t trigger many breach disclosure laws*
    - *Can still obtain high quality analysis outcomes*

**What We Need: Legal Incentives**

- “Notice and Consent” framework discourages application of technological advances
  - We can’t guarantee your privacy, so please allow us to use your data in unsafe ways
  - U.S.: [Enforcement action against Snapchat](https://www.ftc.gov/news-events/press-release/2019/07/ftc-enforces-first-action-against-social-media-platforms) for promising to protect privacy and not doing a good enough job
    - Companies get away with not even trying, as long as they tell you so
- Can legal frameworks acknowledge that privacy is at risk?
  - Require efforts to manage, not eliminate, that risk
Restrictions on Results

- Use of Call Records for Fraud Detection vs. Marketing
  - FCC § 222(c)(1) restricted use of individually identifiable information
  - Until overturned by US Appeals Court
  - 222(d)(2) allows use for fraud detection
- Mortgage Redlining
  - Racial discrimination in home loans prohibited in US
  - Banks drew lines around high risk neighborhoods!!!
  - These were often minority neighborhoods
  - Result: Discrimination (redlining outlawed)
  - What about data mining that “singles out” minorities?

Regulatory Constraints: Use of Results

- Patchwork of Regulations
  - US Telecom (Fraud, not marketing)
    - Federal Communications Commission rules
    - Rooted in antitrust law
  - US Mortgage “redlining”
    - Financial regulations
    - Comes from civil rights legislation
- Evaluate on a per-project basis
  - Domain experts should know the rules
  - You’ll need the domain experts anyway – ask the right questions
Fair Information Practices

1. Notice/Awareness
2. Choice/Consent
3. Access/Participation
4. Integrity/Security
5. Enforcement/Redress
   – Self-Regulation
   – Private Remedies
   – Government Enforcement

http://www.ftc.gov/reports/privacy3/fairinfo.shtm