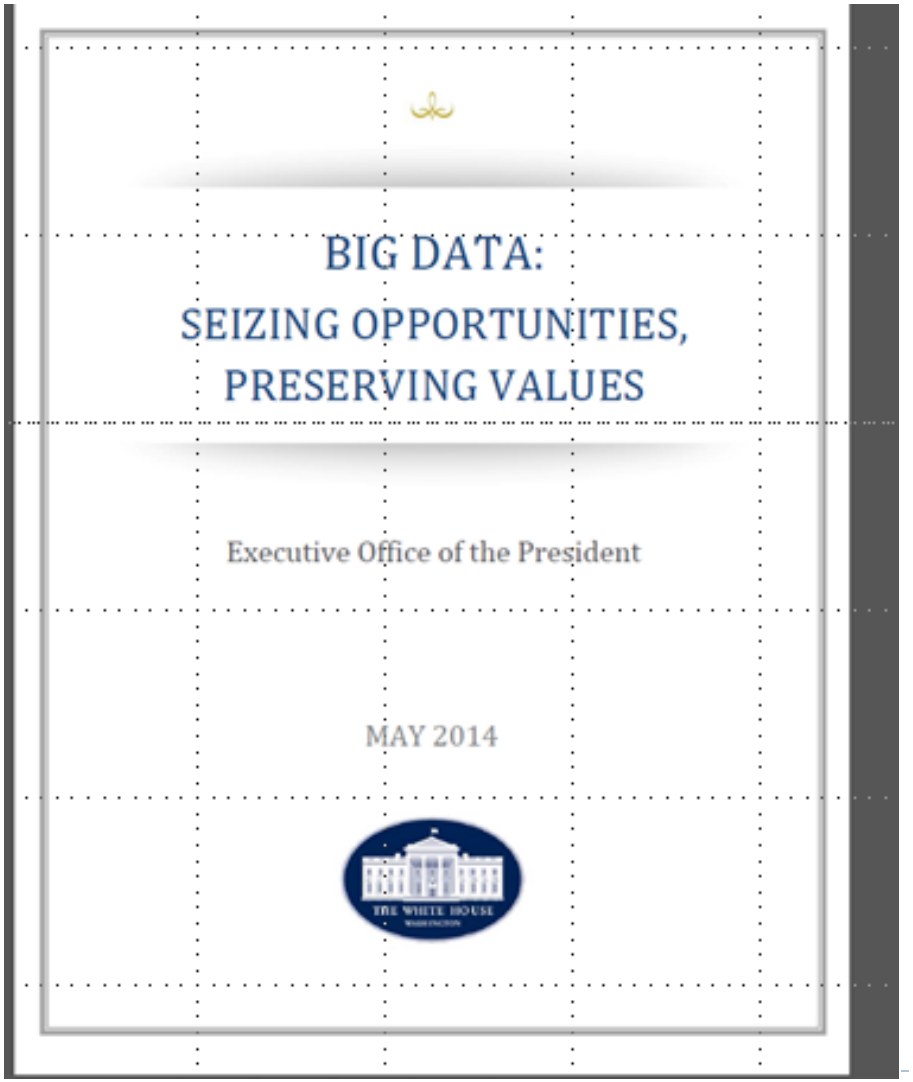
Theory for Society

Cynthia Dwork, Microsoft Research

Podesta Report



Technology Can Erode Values

- Privacy
  - Geolocation data; Call Data Records
  - Smartgrid data
  - *statistics* from pooled genomics data
- Fairness
  - Potholes
  - Consumer Scoring Functions
  - Black-Sounding Names
  - Jobs via linked-in: no applicant pool
- Community of Discourse
  - Narrow-casting, “siloization”

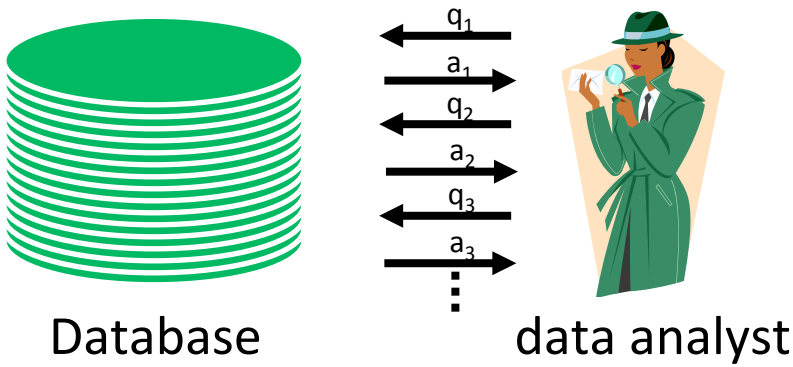
A Role for Theory

- Systems are being built. Promises are made.
  - "Your privacy is important to us."
- *What are they talking about?*
- The answer is literally codified in the system. What are the consequences to the values of interest? What is the boundary between possible and impossible? What can we hope to achieve?

Statistical Data Analytics

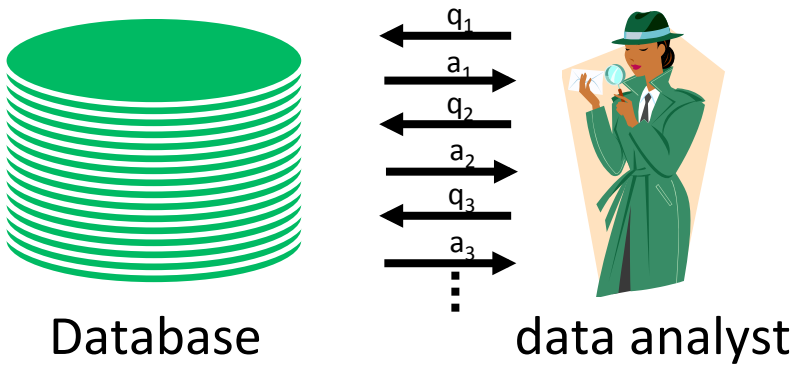
A Surprising Avenue of Privacy Loss

“Just” Statistics?



Tracing Attacks

“Just” Statistics?



Fundamental Law of Info Recovery

- "Overly accurate" estimates of "too many" statistics is blatantly non-private.

Fundamental Law of Info Recovery

- Privacy has a price.

Differential Privacy

- The outcome of any analysis is essentially equally likely, independent of whether any individual joins, or refrains from joining, the dataset.

Differential Privacy [D.,McSherry,Nissim,Smith '06]

M gives ϵ-differential privacy if for all pairs of adjacent data sets x, y, and all events S

$$\Pr[M(x) \in S] \leq e^{\epsilon} \Pr[M(y) \in S]$$

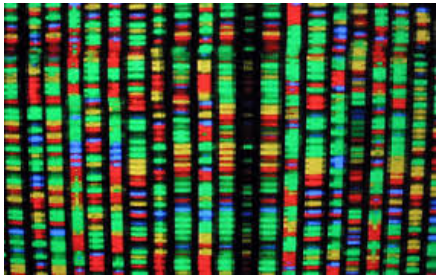
Randomness introduced by M

Differential Privacy [D.,McSherry,Nissim,Smith '06]

M gives ϵ-differential privacy if for all pairs of adjacent data sets x, y, and all events S

$$\Pr[M(x) \in S] \leq e^{\epsilon} \Pr[M(y) \in S]$$

If a bad event is very unlikely when I'm not in dataset (y)
then it is still very unlikely when I am (x)

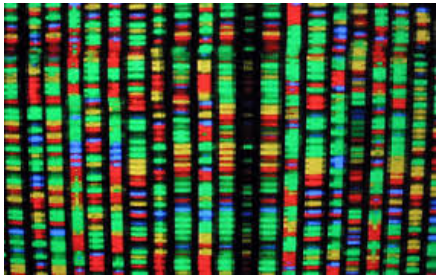


Differential Privacy [D.,McSherry,Nissim,Smith '06]

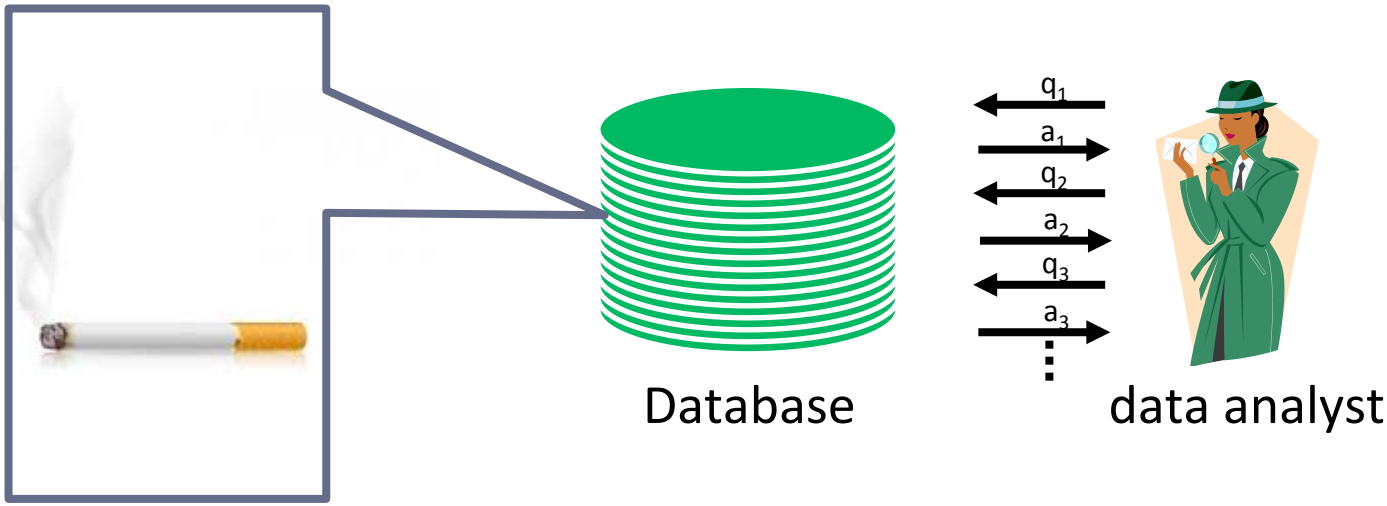
M gives ϵ-differential privacy if for all pairs of adjacent data sets x, y, and all events S

$$\Pr[M(x) \in S] \leq e^{\epsilon} \Pr[M(y) \in S]$$

Extremely strong guarantee. Yet, almost "free": Nearly matches bounds imposed by the Fundamental Law.



Teachings vs Participation



SURGEON GENERAL'S WARNING: Smoking Causes Lung Cancer, Heart Disease, Emphysema, and May Complicate Pregnancy.

Fairness in Classification

Dwork, Hardt, Pitassi, Reingold, Zemel 2012

Algorithms Can Be Biased

- The National Resident Matching Program matches graduating medical students (the applicants) with hospital residency programs.
  - Implementation choice: program-proposing" vs "applicant-proposing
  - Originally program-proposing. Changed in 1997 in response to "a crisis of confidence concerning whether the matching algorithm was unreasonably favorable to employers at the expense of applicants, and whether applicants could 'game the system' by strategically manipulating the [rank order lists] they submitted" (Roth and Peranson, 1997)
  - The new algorithm, adopted in 1997, was also more fair to couples seeking pairs of residencies.

- Classifiers trained on biased or impoverished historical data: loans, college admissions

Forecasts of failure on probation or parole

- “The central question addressed is what role race should play as a predictor when as an empirical matter the majority of perpetrators *and* victims are young, African American males.”
  - R. Berk, The Role of Race in Forecasts of Violent Crime, 2009

Concern: Discrimination

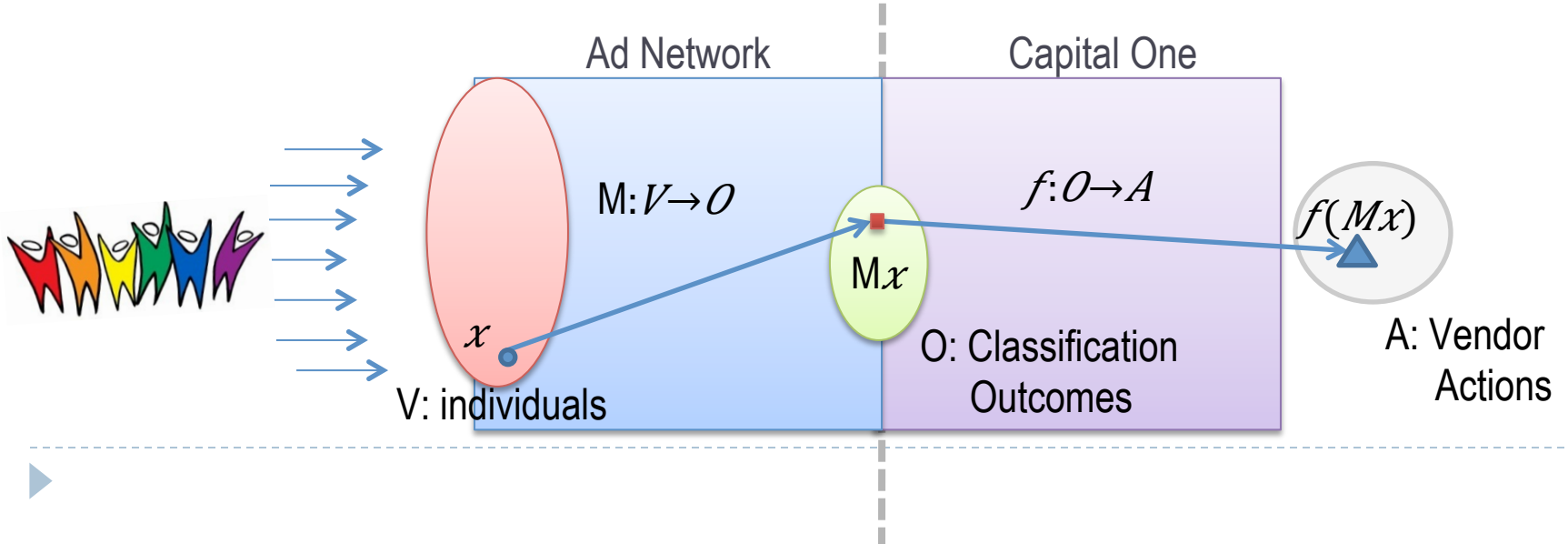
- Population includes minorities
  - Ethnic, religious, medical, geographic
- For this talk: a single protected set S.

Credit Application (WSJ 8/4/10)

User visits capitalone.com

Capital One uses tracking information provided by the tracking network [x+1] to personalize offers

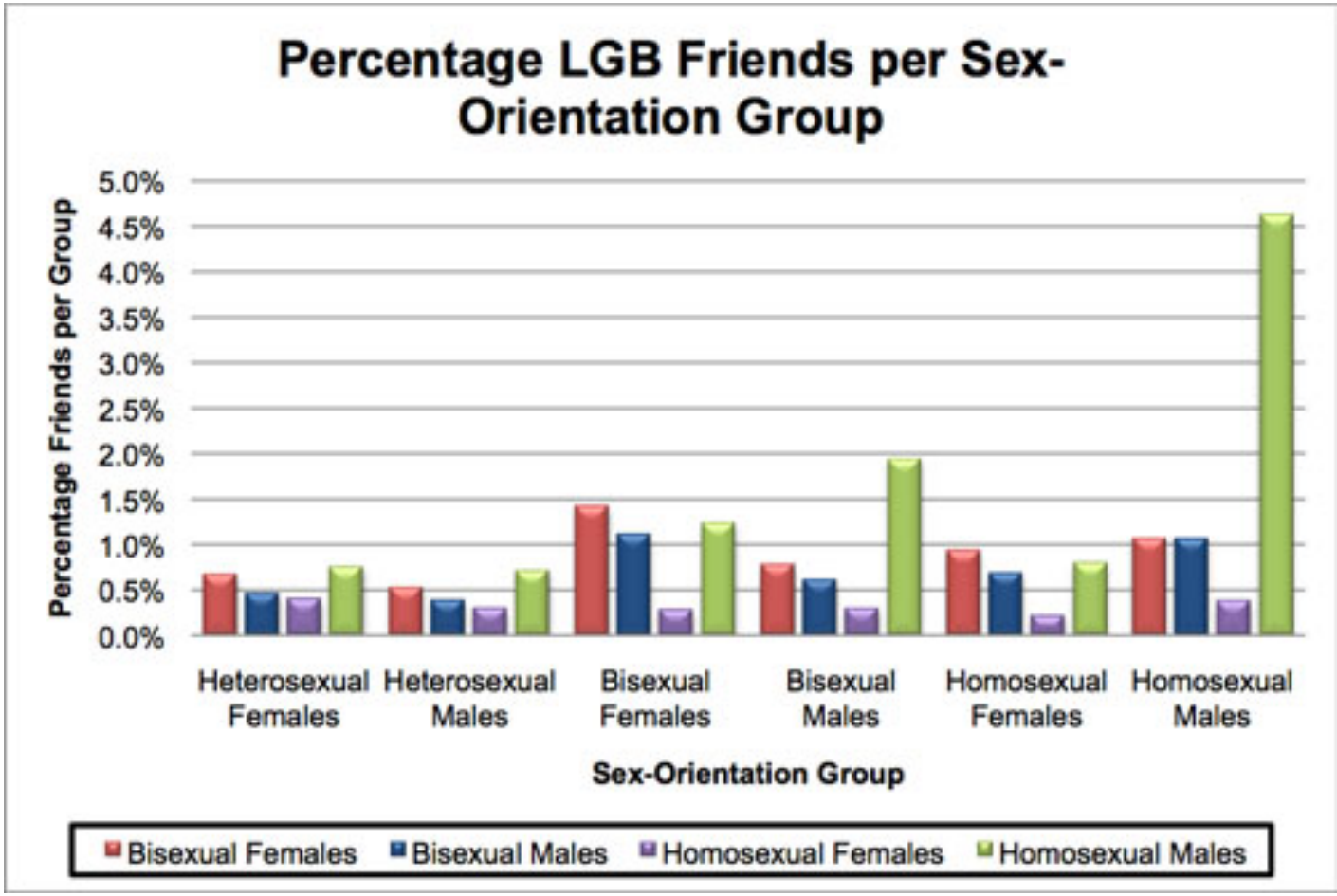
Concern: *Steering* minorities into higher rates (illegal)



Blindness

- Ignore the “membership in S” bit

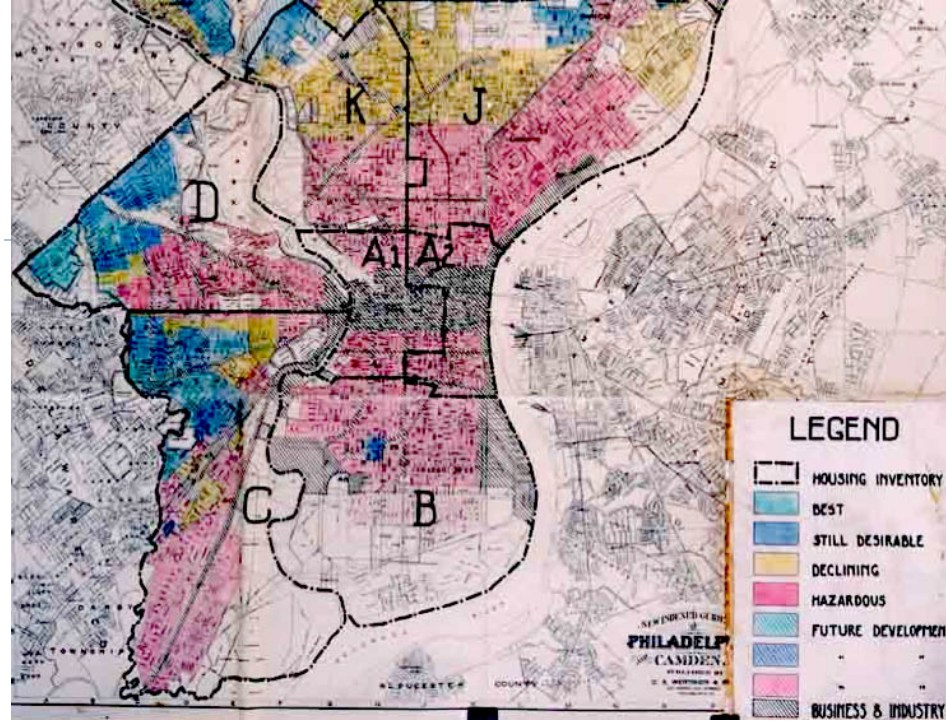
- Membership in S may be holographically embedded in other attributes



Jernigan & Mistree

Design Goals

- Prevent “Catalog of Evils”
  - Redlining, (reverse) tokenism, deliberately targeting “wrong” subset of S,…
- Capture social constraints
  - $\geq$ 1/3 of successful applicants must be female
- Normative approach to preventing discrimination
  - “We do not discriminate, because the system prevents it”
  - …without explicitly finding all redundant encodings
- Retain flexibility for unknown, untrusted, un-auditable user
  - …without needing to “understand” or vet user’s actions

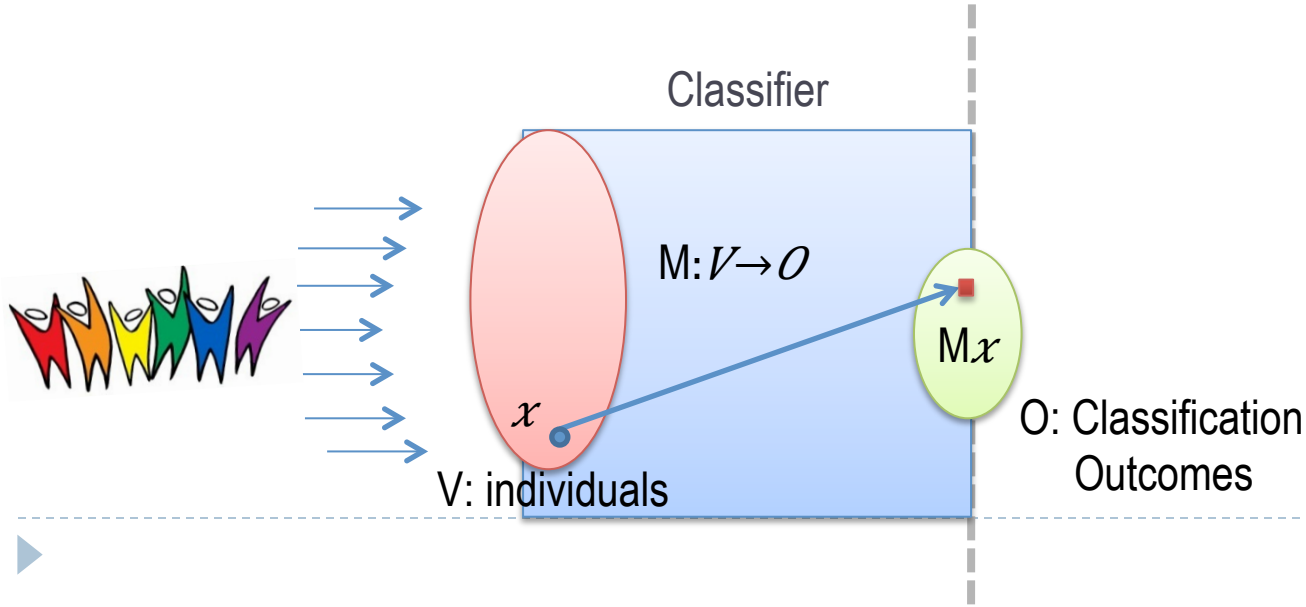


Statistical Parity

- ☑ Demographics of selected group = demographics of population
  - Pr[x in $\mathcal{S}$| outcome = o] = Pr[x in $\mathcal{S}$]
  - Pr[x mapped to o | x in $\mathcal{S}$] = Pr[x mapped to o | x in $\mathcal{S}\uparrow_{\mathcal{C}}$]
  - Completely neutralizes redundant encodings
- ✖ Permits several evils in the catalog
  - E.g., intentionally targeting the subset of $\mathcal{S}$ unable to buy/accept
- ✖ Low utility on $\mathcal{S}$ when multicultural awareness is low
  - Brightest members of $\mathcal{S}$ steered to math, less good to finance
  - Brightest members of $\mathcal{S}\uparrow_{\mathcal{C}}$ steered to finance, less good to math
  - Classifier based on "studying finance" has reduced utility on $\mathcal{S}$
  - Members of $\mathcal{S}$ studying math and science are "close to" members of $\mathcal{S}\uparrow_{\mathcal{C}}$ studying finance, for purposes of "bright student" classifier

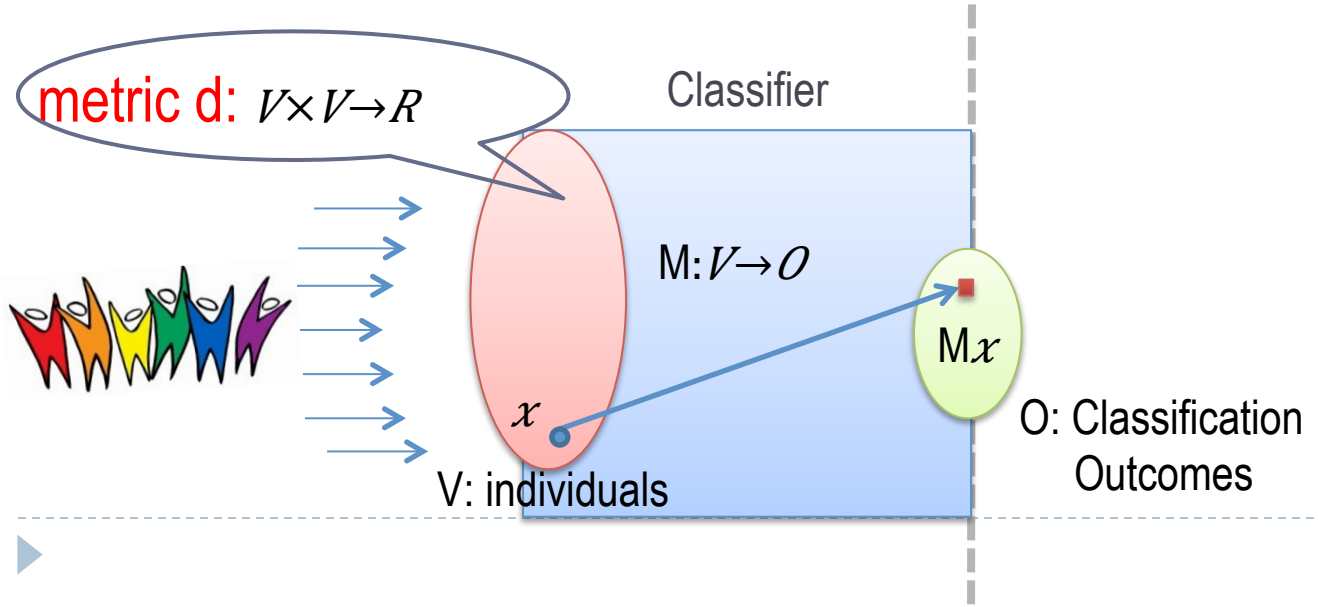
Individual Fairness

- People who are similar with *respect to a specific classification task* should be treated similarly
  - S + math $\sim_{S\uparrow C}$ + finance
  - "Fairness Through Awareness"



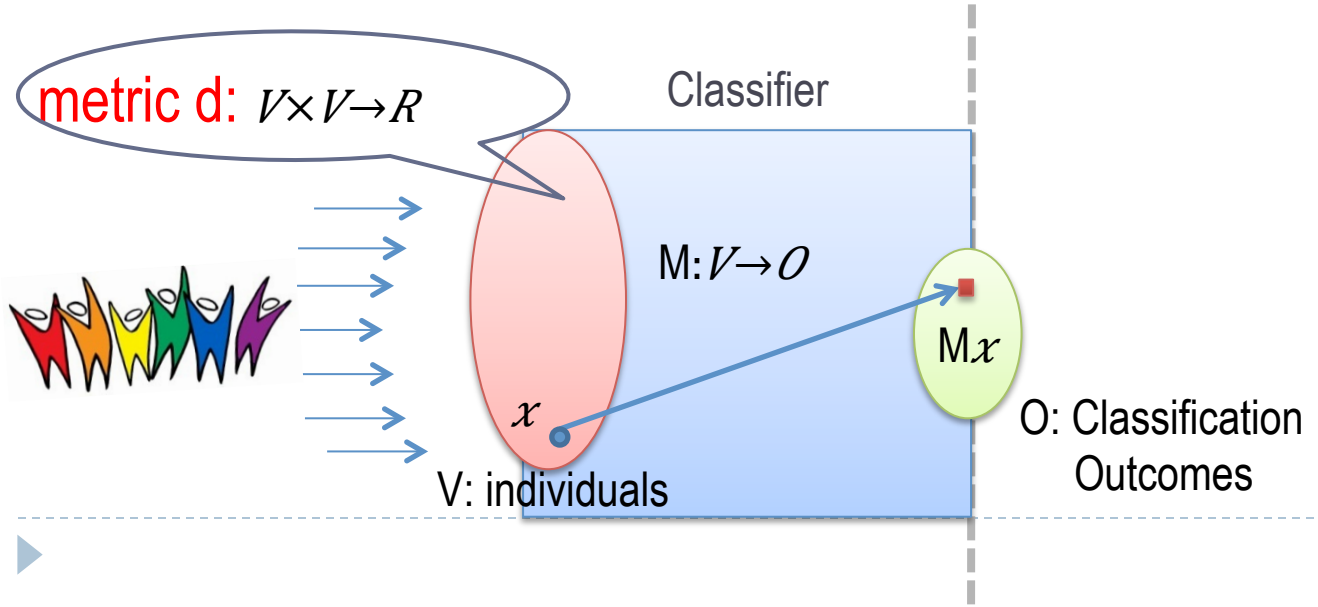
Individual Fairness

- Science Fiction: task-specific similarity metric
  - Examples: credit score; college admissions formulae
  - Sunshine: Public, open to discussion, revision, regulation



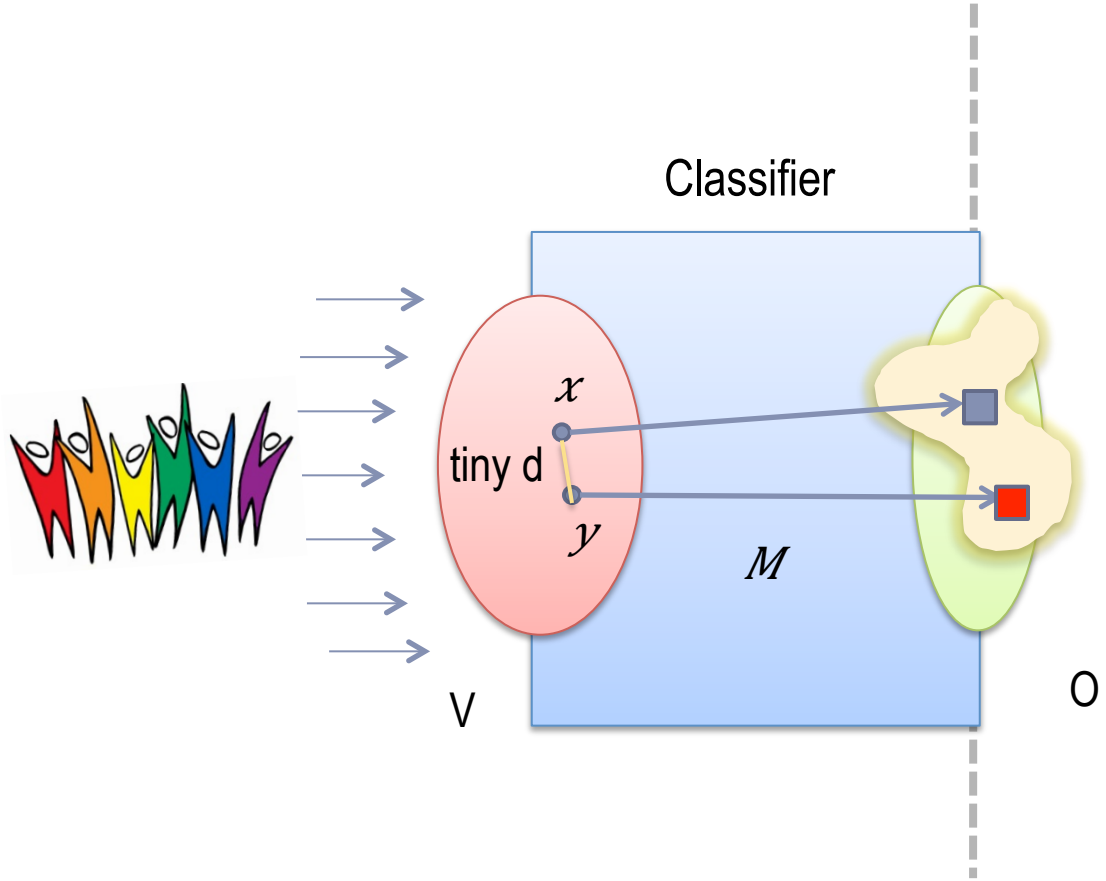
Unarticulated Discrimination

- Sunshine: Public, open to discussion, revision, regulation
  - Criminal record and *The New Jim Crow*
    - Discrimination in housing, employment, public assistance, voting,… is legal
    - Practices in the US "war on drugs" that create the criminal classification are highly racially biased



Michelle Alexander

Individual Fairness and Drawing the Line



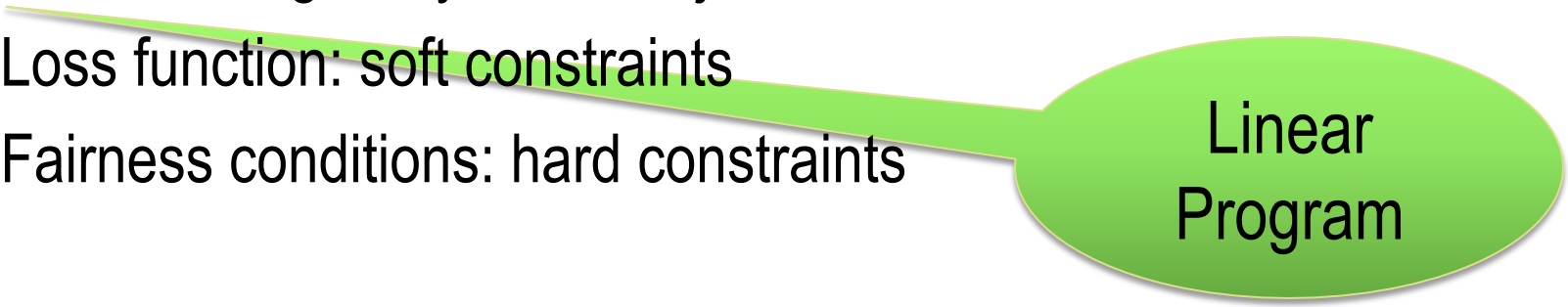
$$M: V \rightarrow \Delta(O)$$
$$||M(x) - M(y)|| \leq d(x,y)$$

Assemble Ingredients

"Minimize user's avg utility loss, subject to the fairness conditions."

Loss function: soft constraints

Fairness conditions: hard constraints



Can modify the approach to obtain fair affirmative action

Affective Computing and Emotional Manipulation

Your Name Here

A REPORTER AT LARGE | JANUARY 19, 2015 ISSUE

WE KNOW HOW YOU FEEL

Computers are learning to read emotion, and the business world can't wait.

BY RAFFI KHATCHADOURIAN

SHARE TWEET

Three years ago, archivists at A.T. & T. stumbled upon a rare fragment of computer history: a short film that Jim Henson produced for Ma Bell, in 1963. Henson had been hired to make the film for a conference that the company was convening to showcase its strengths in machine-to-machine communication. Told to devise a faux-

A REPORTER AT LARGE | JANUARY 19, 2015 ISSUE

WE KNOW HOW YOU FEEL

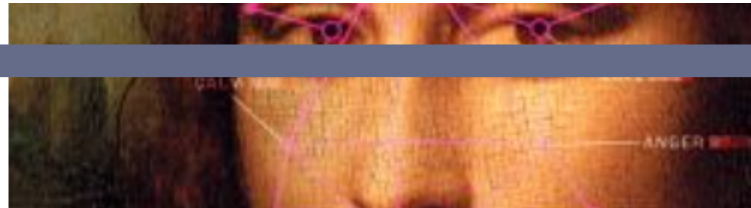
Computers are learning to read emotion, and the business world can't wait.

BY RAFFI KHATCHADOURIAN

SHARE TWEET

Experts on the voice have trained computers to identify deep patterns in vocal pitch, rhythm, and intensity; their software can scan a conversation between a woman and a child and determine if the woman is a mother, whether she is looking the child in the eye, whether she is angry o frustrated or joyful. Other machines can measure sentiment by assessing the arrangement of our words, or by reading our gestures. Still others can do so from facial expressions.

company was convening to showcase its



strengths in machine-to-machine communication. Told to devise a faux

Facebook is Not My Friend

TECH

Furor Erupts Over Facebook's Experiment on Users

Almost 700,000 Unwitting Subjects Had Their Feeds Altered to Gauge Effect on Emotion

http://www.wsj.com/articles/furor-erupts-over-facebook-experiment-on-users-1404085840

My Advertiser is Not My Friend

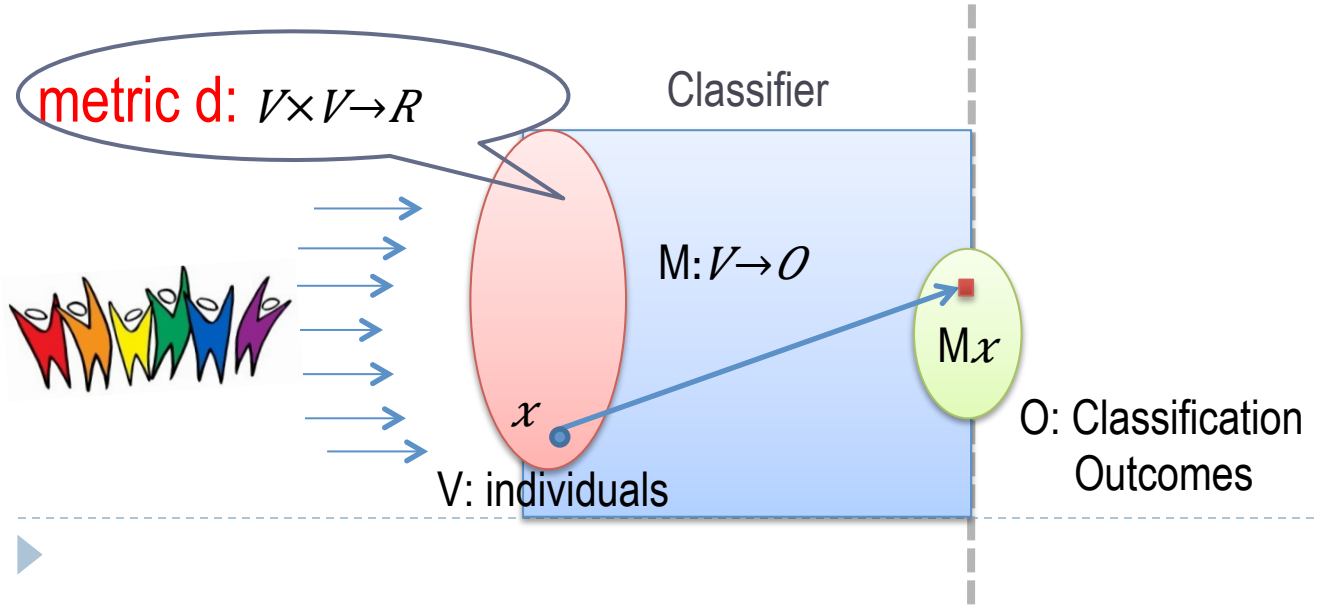
- Creates demand
- Doesn't have my interests at heart
  - If I am sad, my advertiser will suggest I buy …(chocolate?)
- Does my advertiser know why I am sad?
  - Maybe sad is appropriate
- Does my advertiser want to keep me sad?
- Exploiting my emotional state for financial gain
- Manipulating my emotional state for continued financial gain

Thank you!

Snowbird, July 18, 2016

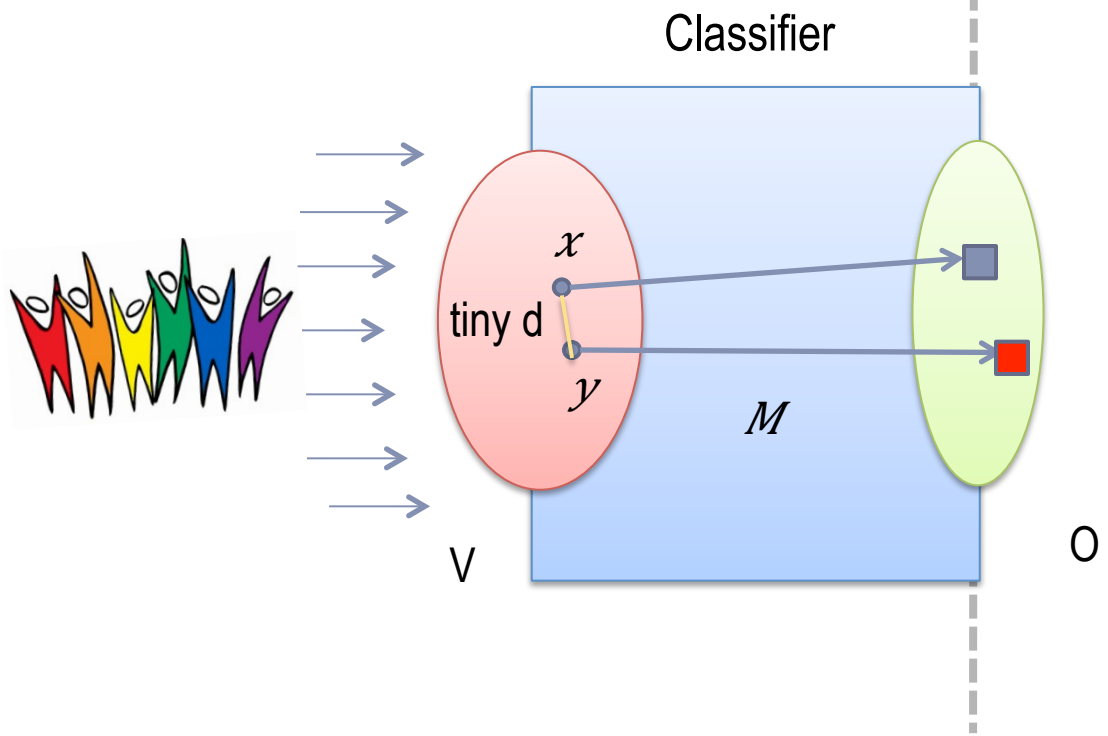
Individual Fairness

- Science Fiction: task-specific similarity metric
  - Ideally, ground truth (hence "fair")
  - In reality, society's "best approximation"
    - We don't attempt to resolve the philosophical questions



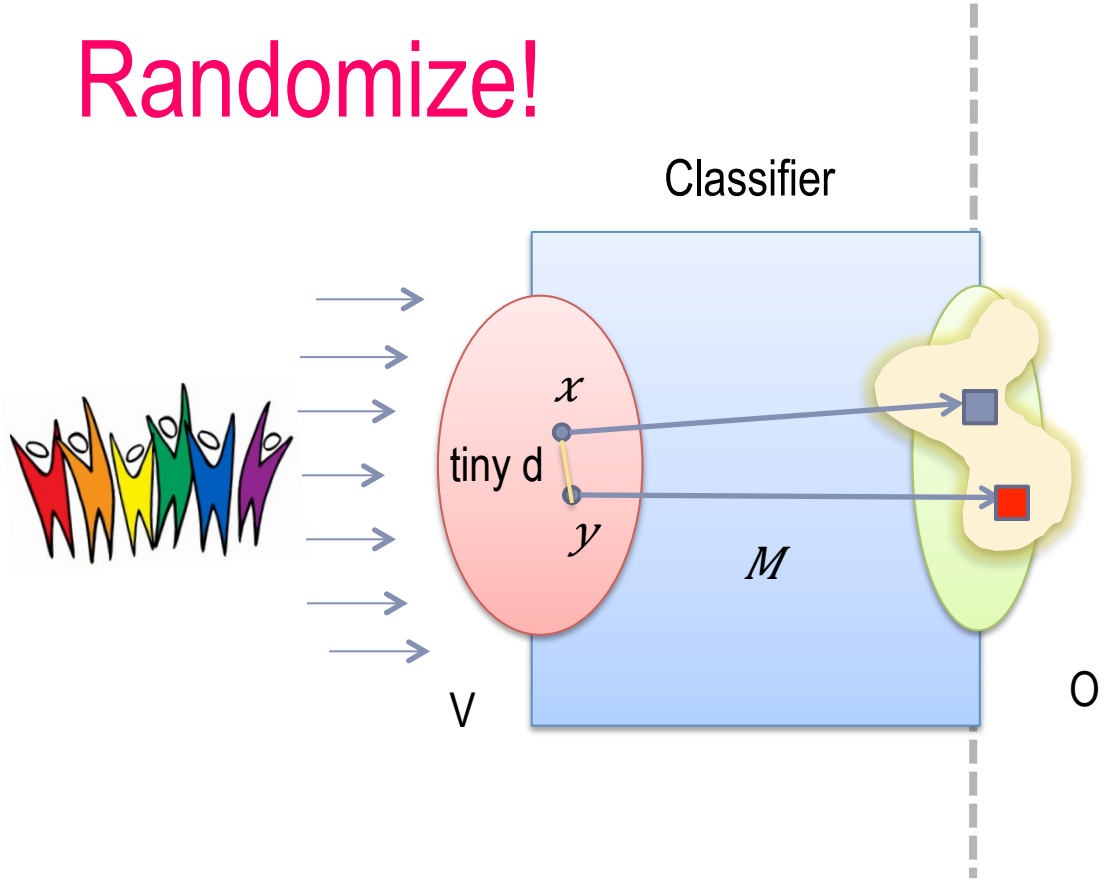
Individual Fairness

Meaningful?



Individual Fairness

Randomize!



$M: V \to \Delta(O)$

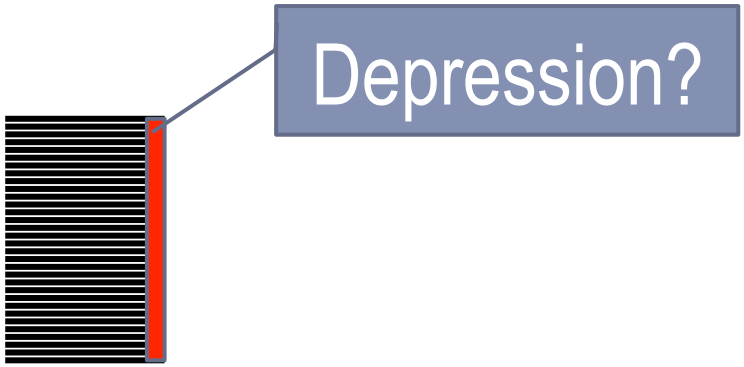
Unarticulated Discrimination

- Sunshine: Public, open to discussion, revision, regulation
  - Criminal record and *The New Jim Crow*
    - Discrimination in housing, employment, public assistance, voting,… is legal
    - Practices in the US "war on drugs" that create the criminal class are highly racially biased
  - "The basis of the decision to single out particular passengers during a suspicionless sweep is less likely to be inarticulable than unspeakable."

Justice Thurgood Marshall, *Florida vs Bostick*

The Statistics Masquerade

- Differencing Attack
  - *How many members of House of Representatives have sickle cell trait?*
  - *How many members of House, other than the Speaker, have the trait?*
- Needle in a Haystack
  - Determine presence of an individual's genomic data in GWAS case group
- The Big Bang attack
  - Reconstruct "depression" bit column



Dinur and Nissim'03 Homer+'08