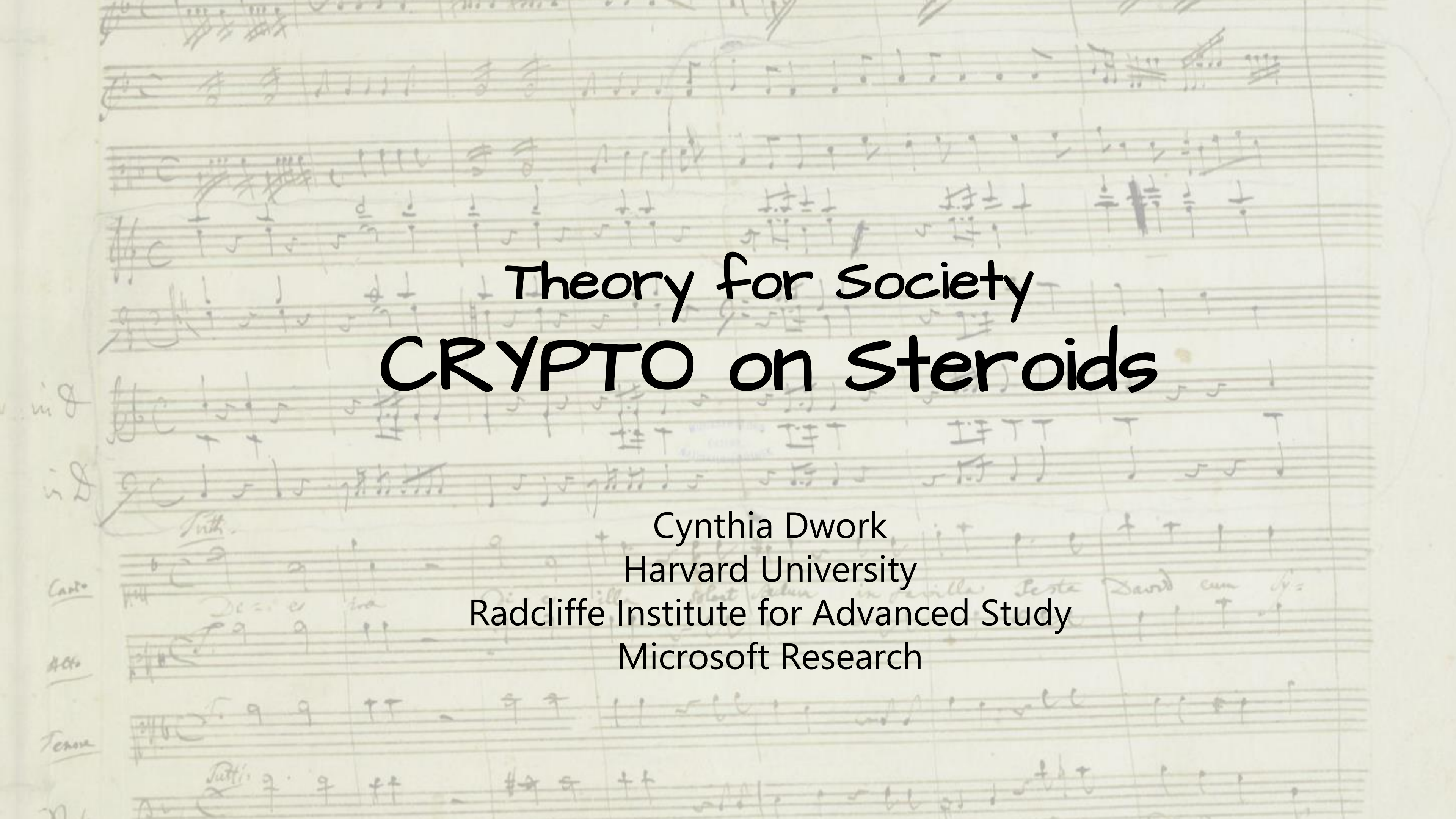


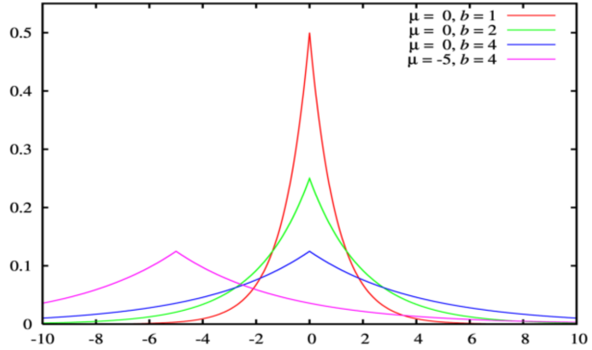
Theory for Society

CRYPTO on Steroids

Cynthia Dwork
Harvard University
Radcliffe Institute for Advanced Study
Microsoft Research

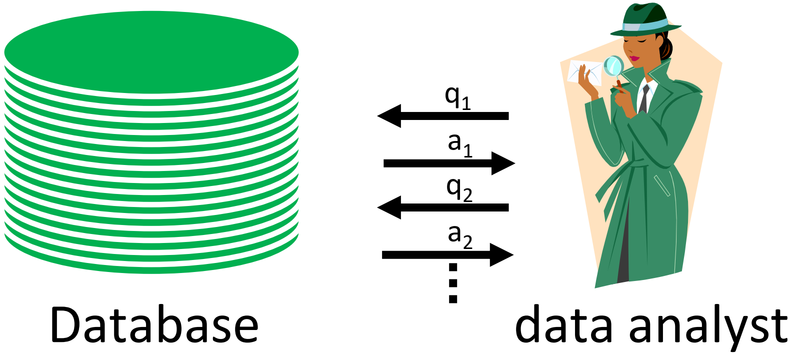
Motivation

- The adversary is ubiquitous
- The adversary is Byzantine



Differential Privacy

Model for this Talk



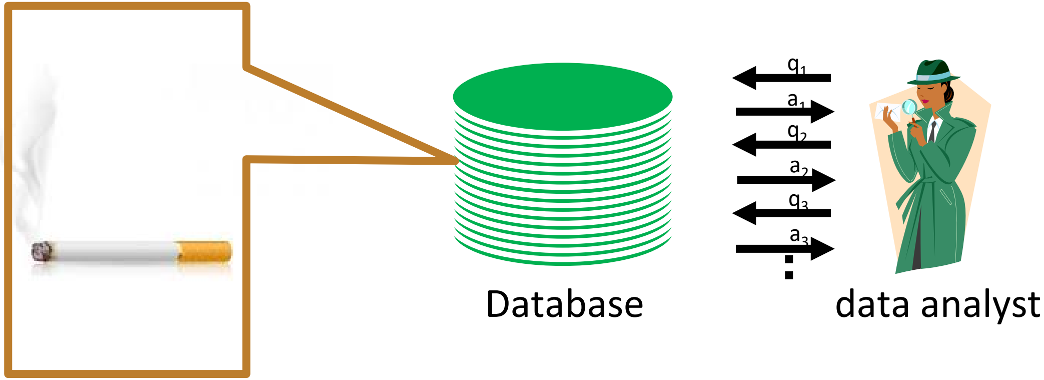
Other models: “local”, “federated”

Differentially Private Data Analysis

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

- “learn the same things” whether or not I am in the database

Teachings vs Participation



q_1

a_1

q_2

a_2

q_3

a_3

Database

data analyst

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

Differentially Private Data Analysis

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

- "learn the same things" whether or not I am in the database
- Privacy and generalizability are aligned!

Differential Privacy

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

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

"Privacy Loss"

Randomness introduced by M

Bounded Ratio

Key Properties

- Future-Proof
  - Resilient to present – or future – auxiliary information

- Group Privacy
  - Automatically yields $k\epsilon$-differential privacy for groups of size k

- Composes Gracefully and Automatically
  - Differential privacy is programmable

Beyond CRYPTO

Addresses privacy problems arising when everything is working properly

- ○ How many members of the House of Representatives are cooperating?
- ○ How many members of the House other than the Speaker are under investigation?

Privacy Loss

Fix adjacent x, y, draw $c \leftarrow M(x)$

$$\text{PrivacyLoss}_{x,y}(c) = \ln\left[\frac{\Pr[M(x) = c]}{\Pr[M(y) = \mathrm{c}]}\right]$$

- Can be positive, negative (or infinite)
- In pure ϵ-DP it is always bounded by ϵ

Privacy Loss is a Random Variable

- Even more powerful composition results
- Variants of DP with accuracy matching the lower bounds of the Fundamental Theorem of Data Recovery

The Exponential Mechanism

- $f(x) \in \{\xi_1, \xi_2, \dots, \xi_k\}$
  - Strings, experts, small databases, prices…
  - Each ξ has a utility for x, denoted $u(x,\xi)$
  - $\Delta u = \max_{\xi,\mathrm{adj}(x,y)} |u(x,\xi) - u(y,\xi)|$
- Intuition: Output ξ with probability $\propto e^{u(x,\xi)\epsilon/\Delta \mathrm{u}}$

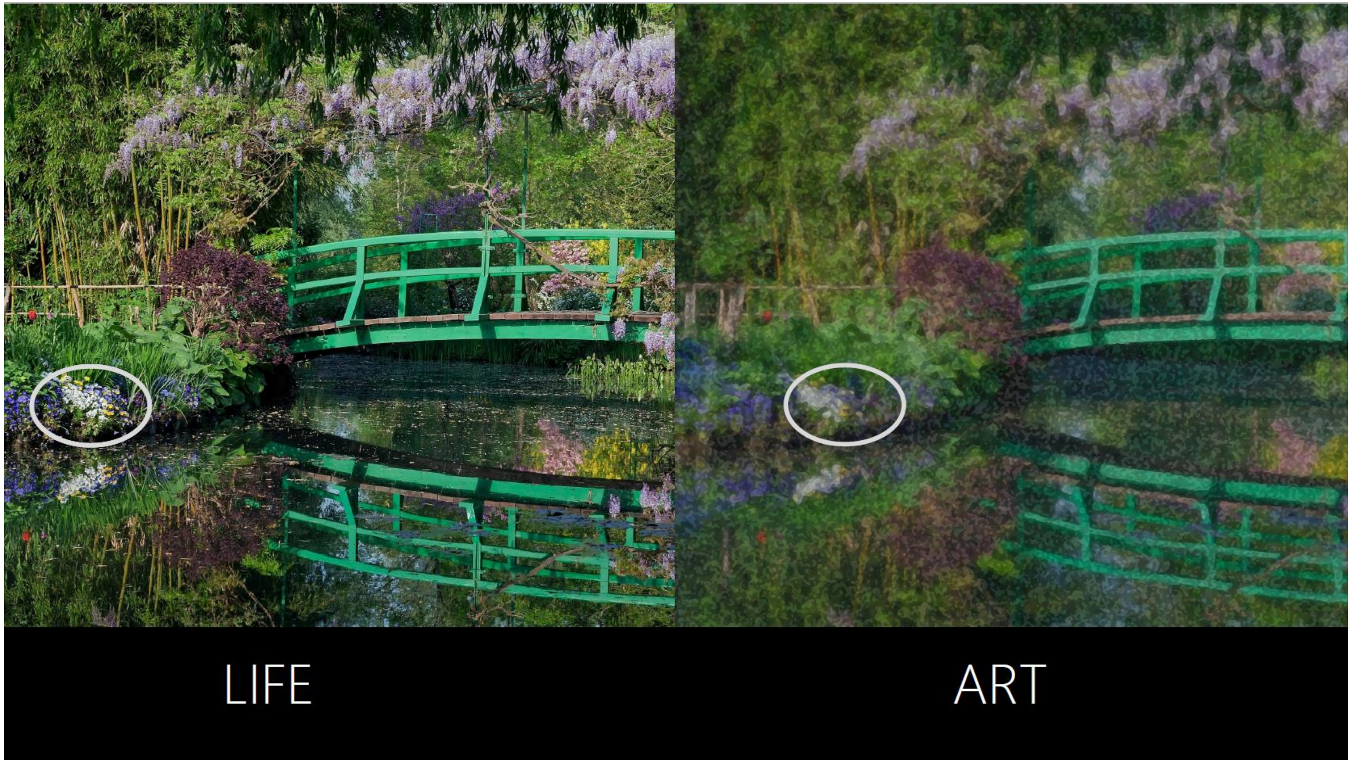
$$\left[\frac{\exp(u(x,\xi))}{\exp(u(y,\xi))}\right]^{\epsilon/\Delta u} = \left[e^{u(x,\xi)-u(y,\xi)}\right]^{\epsilon/\Delta u} \le e^{\epsilon}$$

[McSherry,Talwar 2007]

Rich Algorithmic Literature

- Counts, linear queries, histograms, contingency tables (marginals)
- Location and spread (eg, median, interquartile range)
- Dimension reduction (PCA, SVD), clustering
- Support Vector Machines
- Sparse regression/LASSO, logistic and linear regression
- Boosting, Multiplicative Weights
- Combinatorial optimization, mechanism design
- Privacy Under Continual Observation, Pan-Privacy
- Finite sample confidence intervals
- Statistical Queries learning model, PAC learning
- False Discovery Rate control in multiple hypothesis testing
- (Stochastic) gradient descent, deep learning
- …
- *The Algorithmic Foundations of Differential Privacy,* Dwork and Roth, August 2014

Which is "Truth"?



Open access, freely av

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

is characteristic of the fiel vary a lot depending on w field targets highly likely r or searches for only one c

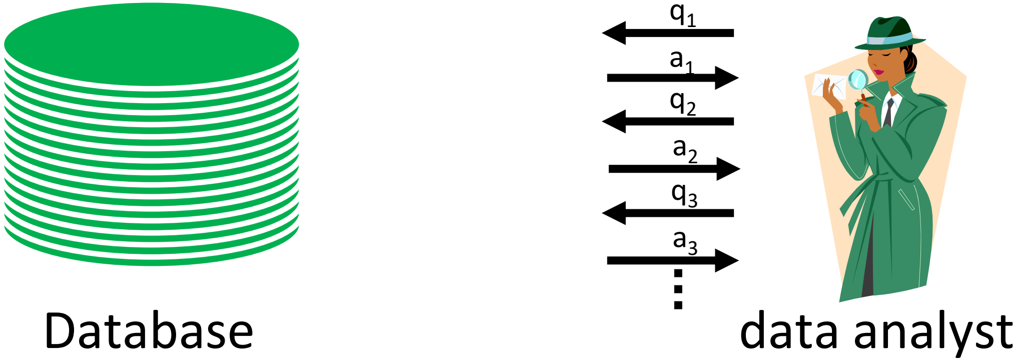
Adaptive Data Analysis

Adaptivity Arises Naturally

- Natural learning procedures (like gradient descent) adaptively query the data.

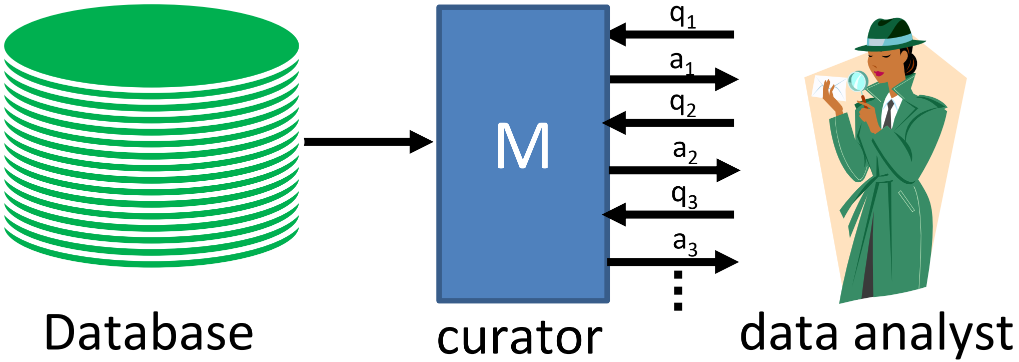
- More insidious: studies conducted by researchers who have read papers that used the same data set must be considered adaptive.

Recall: DP Holds Under Adaptive Composition



- q_i depends on $a_1, a_2, \ldots, a_{i-1}$

Differential Privacy Addresses Adaptivity



- If M ensures differential privacy, then *the choice of* q_i can't reveal "too much" about the database
- From "one-shot" generalization to "generalization under composition"

[Dwork, Feldman, Hardt, Pitassi, Reingold, Roth '14]

Intuition

Fix a query, eg, “What fraction of the population is over 6 feet tall?”
Almost all large data sets will give an approximately correct reply

- Most data sets are representative with respect to this query

If in the process of adaptive exploration, the analyst finds a query for which the training set is not representative, then she “learned something significant” about the set.

- Preserving the “privacy” of the data may prevent over-fitting.

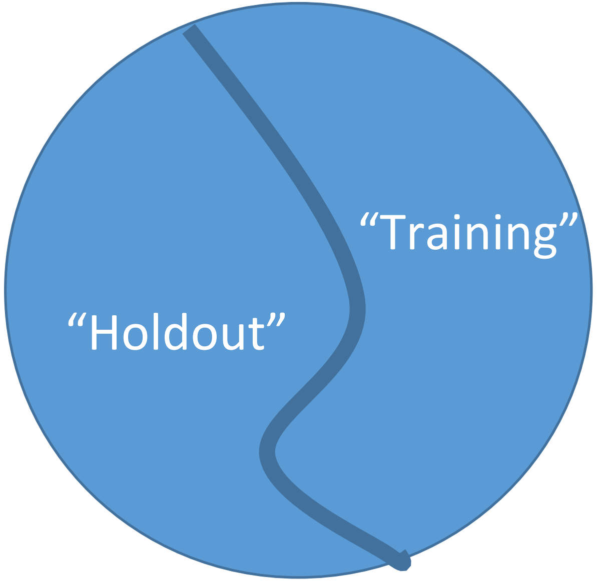
Max Information

- For jointly distributed (D, Φ), define $I_\infty(D; \Phi)$ as min k for which $\forall d$ in support of D, $\forall \phi$ in support of Φ,
$$\Pr[D = d | \Phi = \phi] \leq 2^k \Pr[D = d]$$
  - D will be the data set random variable
  - Φ will be the query random variable chosen by the analyst interacting with $d \sim D$
- Fact: $I_\infty(D; \Phi) = I_\infty(\Phi, D)$
- DP bounds max information $(I_\infty(\Phi, D))$
  - It's not the only way, but it is a way that has huge algorithmic support.
- Bounding max information $(I_\infty(D; \Phi))$ ensures d remains representative
  - The query does not say much about the dataset; generalization persists
- Repeat!

Closure under post-processing

Composition

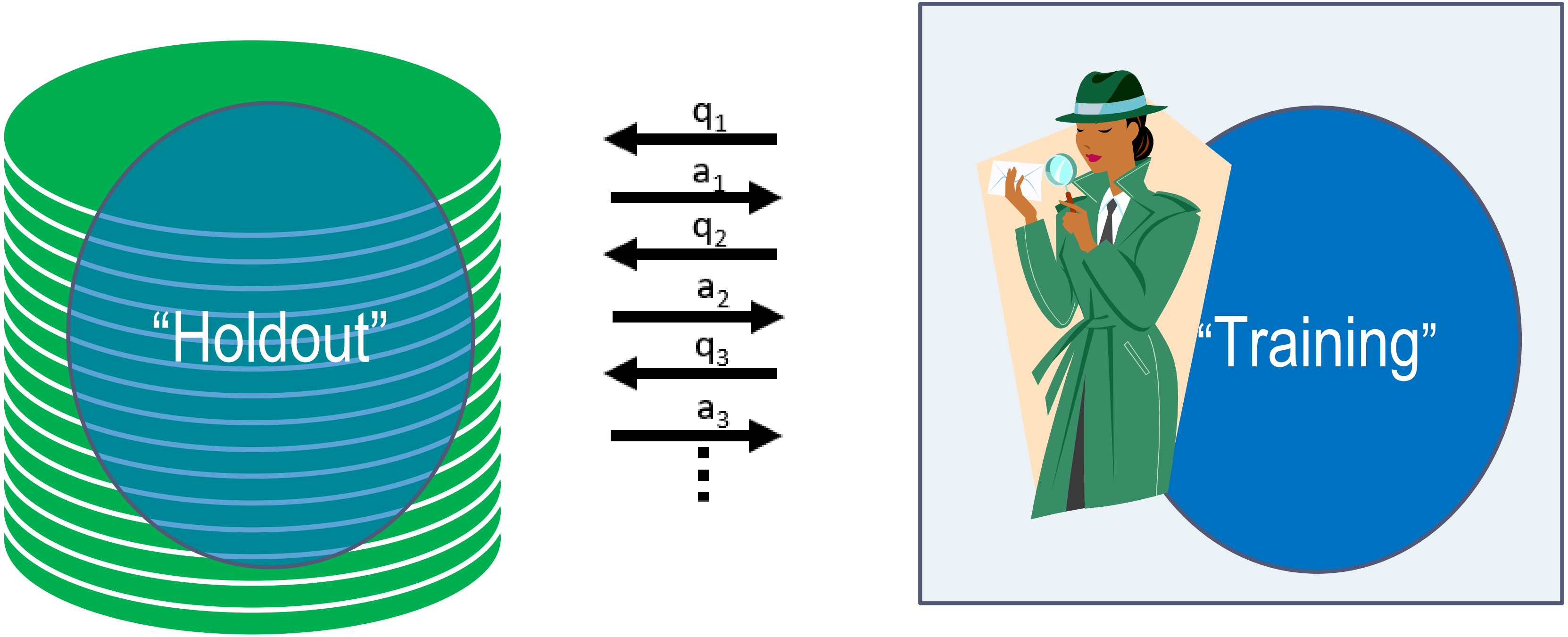
The Re-Usable Holdout



- Learn on the training set
- Check against holdout via a differentially private mechanism
- Future exploration does not significantly depend on *H*
  - *H stays fresh!*

[Dwork, Feldman, Hardt, Pitassi, Reingold, Roth ’14]

Conceptually



Algorithmic Fairness

Population is diverse: ethnic, religious, geographic, medical, gender, class, sexual preference, etc.

- Bank receives detailed user information before serving a page
- Concern: steering minorities to credit card offerings of less desirable terms

Algorithmic Fairness

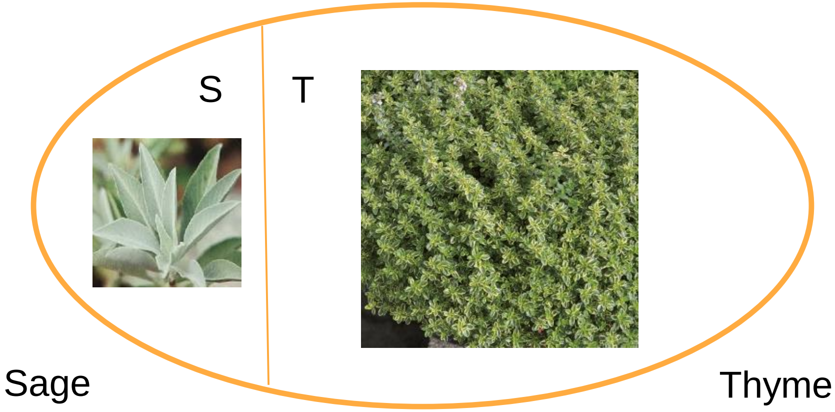
Population is diverse: ethnic, religious, geographic, medical, gender, class, sexual preference, etc.

- “Hiding” the sensitive information does not work

Algorithmic Fairness

Population is diverse: ethnic, religious, geographic, medical, gender, class, sexual preference, etc.

- “Hiding” the sensitive information does not work
- Culturally aware algorithms are more accurate



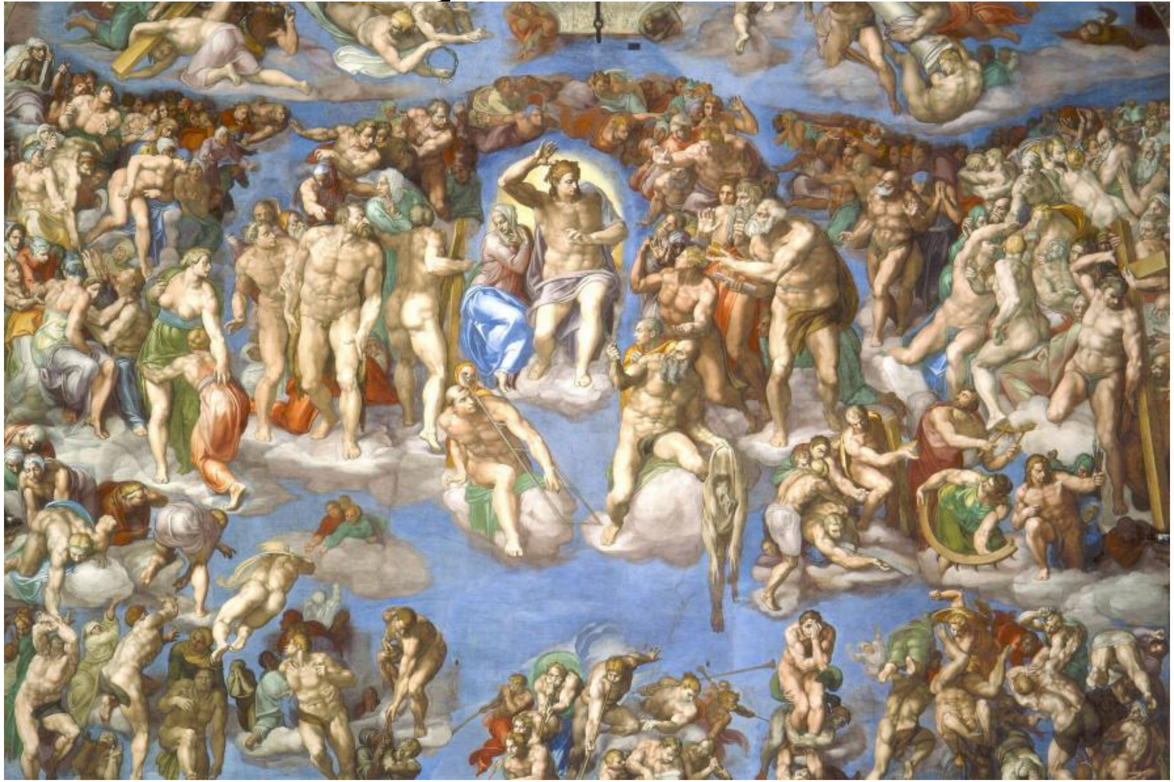
Algorithmic Fairness

Population is diverse: ethnic, religious, geographic, medical, gender, class, sexual preference, etc.

- "Train" on historical data?
    - Imbibe biases in the data
- *No general source of truth*

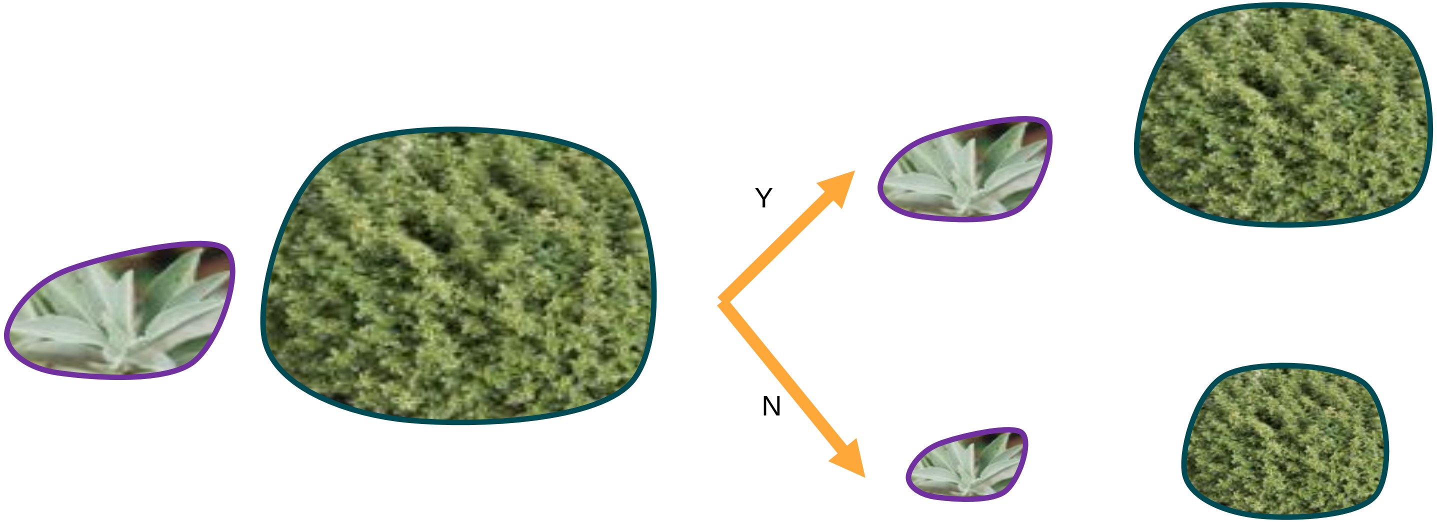
Defining Fairness

Classification Algorithms



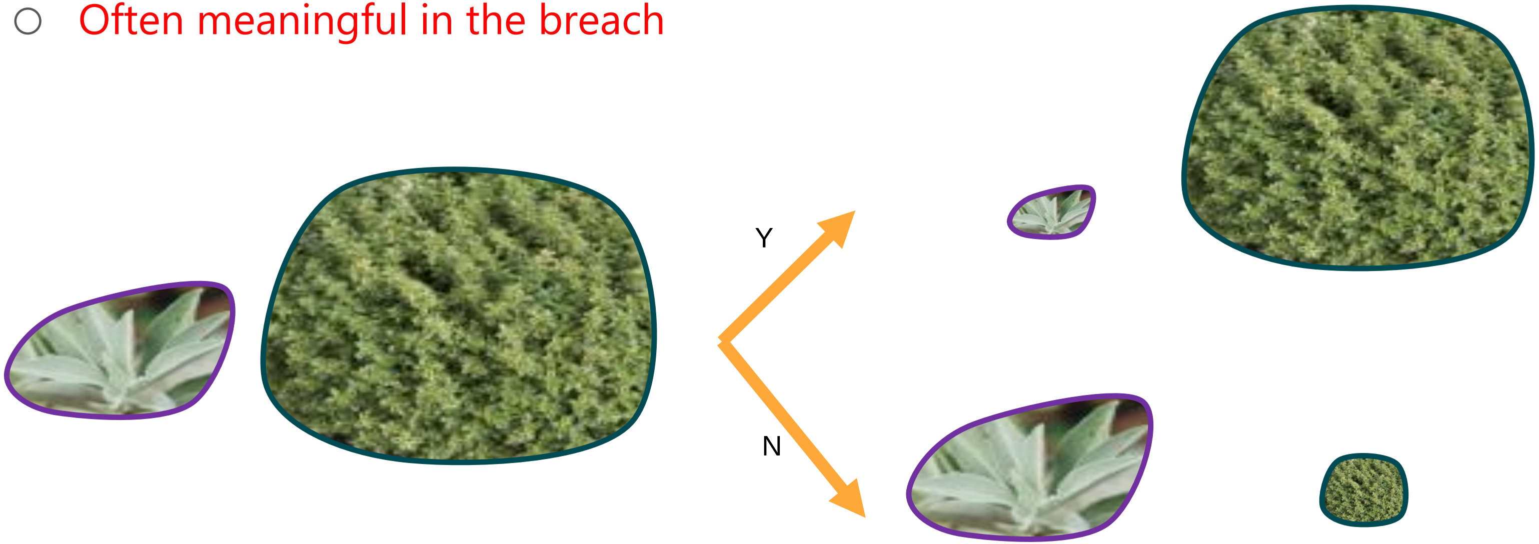
Defining Fairness for Groups

- Group fairness properties are statistical requirements
    - Statistical Parity: demographics of people with positive (negative) classification are the same as the demographics of the general population



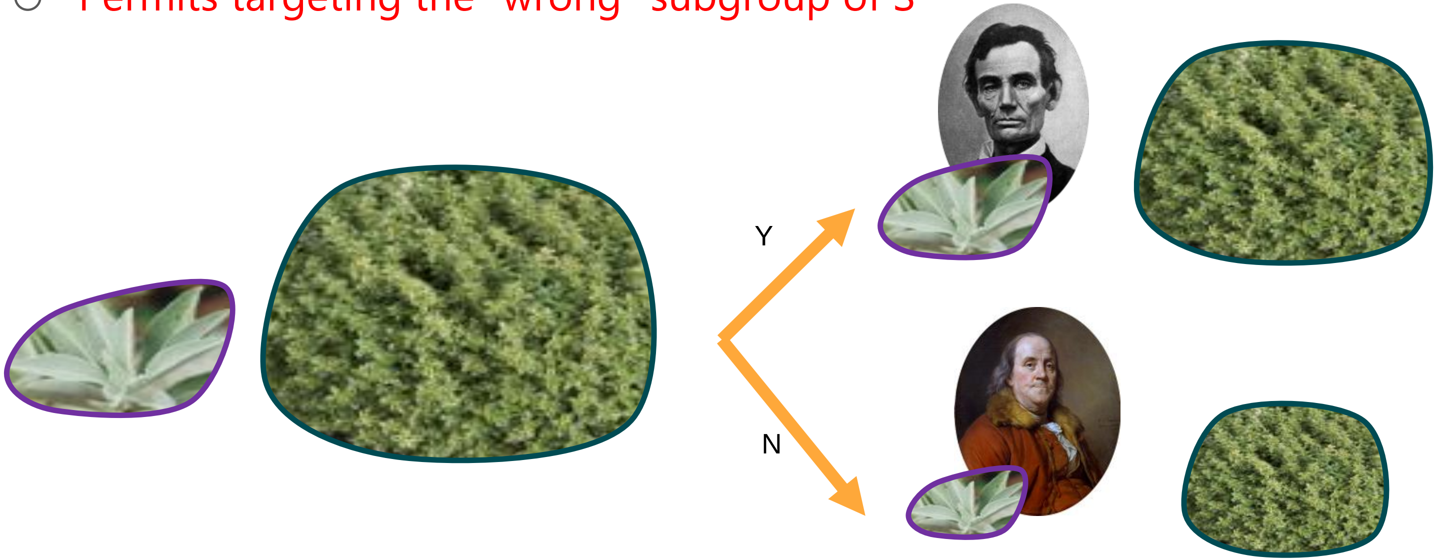
Defining Fairness for Groups

- Group fairness properties are statistical requirements
  - Statistical Parity: demographics of people with positive (negative) classification are the same as the demographics of the general population
  - Often meaningful in the breach



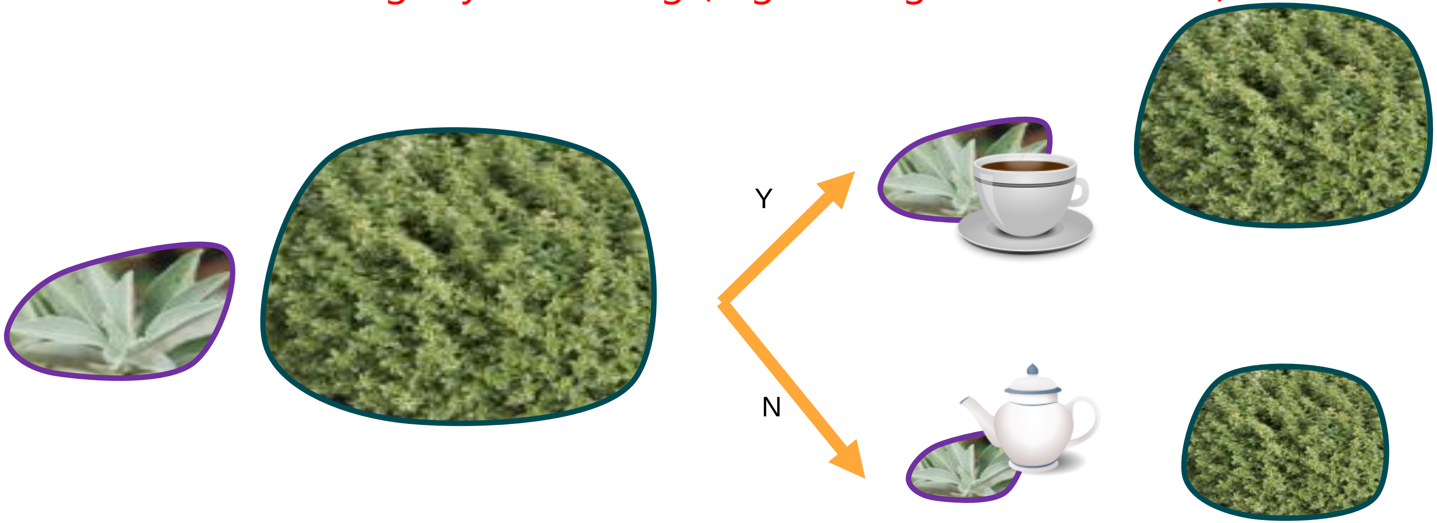
Defining Fairness for Groups

- Group fairness properties are statistical requirements
  - Statistical Parity: demographics of people with positive (negative) classification are the same as the demographics of the general population
  - Permits targeting the “wrong” subgroup of S



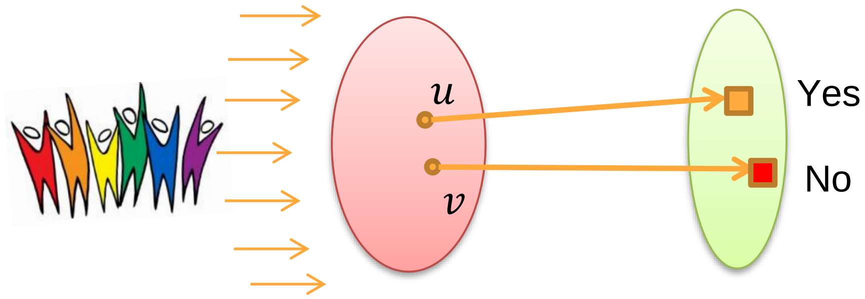
Defining Fairness for Groups

- Group fairness properties are statistical requirements
  - Statistical Parity: demographics of people with positive (negative) classification are the same as the demographics of the general population
  - Permits fairness gerrymandering (sage eating coffee drinkers)



Individual Fairness

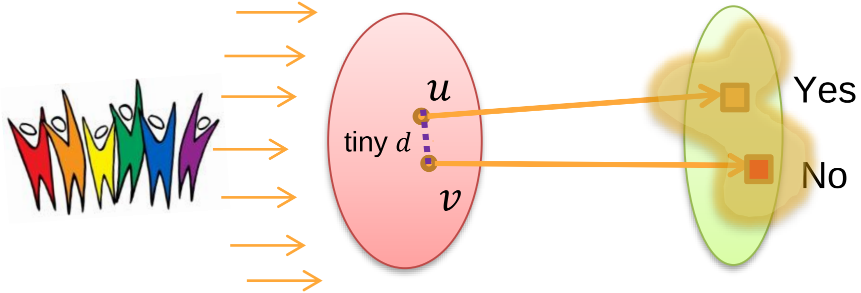
"Similar people" are treated similarly



Need "right" notion of (dis)similarity $d(u, v)$ for the specific classification task

Individual Fairness

“Similar people” have similar *probabilities* of “Yes” and “No” outcomes



$$C: U \rightarrow \Delta(O)$$
$$\left|\left|C(x) - C(y)\right|\right| \leq d(x, y)$$

Dwork, Hardt, Pitassi, Reingold, Zemel 2012

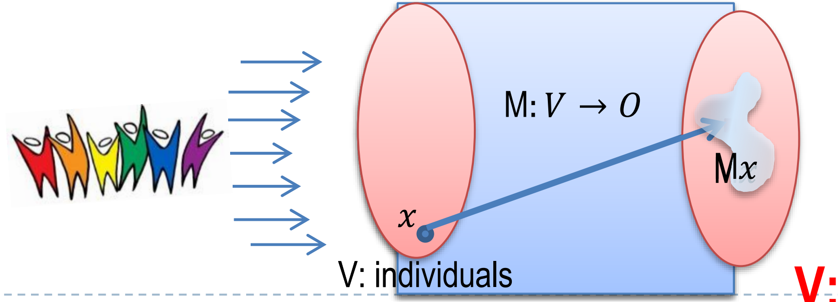
Lipschitz Mappings

	Differential Privacy	Individual Fairness
Objects	Databases	Individuals
Outcomes	Output of statistical analysis	Classification outcome
Similarity	General purpose metric	Task-specific metric

- Exciting possibility: use dp techniques for fairness?
  - Yes, we can!
  - **Theorem:** Exponential mechanism of [MT07] yields individual fairness and small loss in bounded doubling dimension.

Using DP Techniques for Fairness

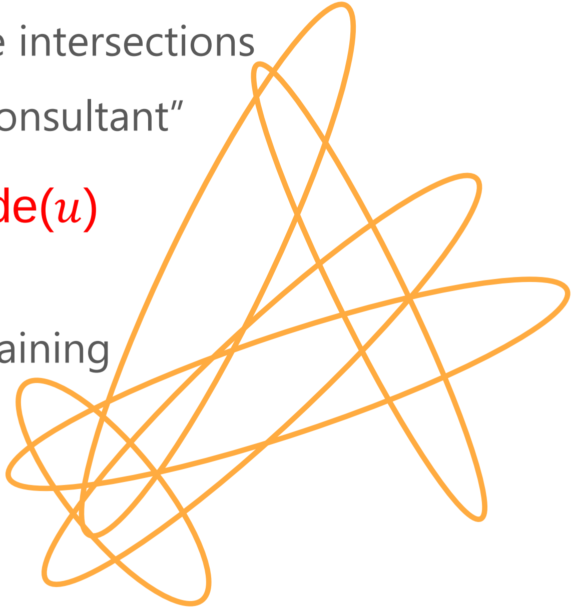
- Set $O = V$ (map individuals to distributions over individuals)
- "Smear" individual across neighbors
  - Probability of mapping v to v' is proportional to $e^{-d(v,v')}$
  - Fairness: Close neighbors mapped to similar distributions on individuals
  - Small loss: Under suitable conditions, expect to be mapped to near neighbor



V: individuals

Veins of Algorithmic Work

- “Standard” optimization techniques when a metric is known
- “Exploration vs exploitation” techniques in bandit settings
- Fairness/calibration for (very) large numbers of large, overlapping, groups
  - Addresses “intersectional” fairness/calibration for large intersections
  - Fairness results leverage limited access to a “fairness consultant”
- Learn a “fair” representation $u \Rightarrow \text{Encode}(u)$
  - Censors sensitive information
  - Keeps enough other information to permit standard training



Dwork, Hardt, Pitassi, Reingold, Zemel Joseph, Kearns, Morgenstern, Roth
Kearns, Neel, Roth, Wu Hebert-Johnson, Kim, Reingold, Rothblum.
Zemel, Wu, Swersky, Pitassi, Dwork Edwards and Storkey

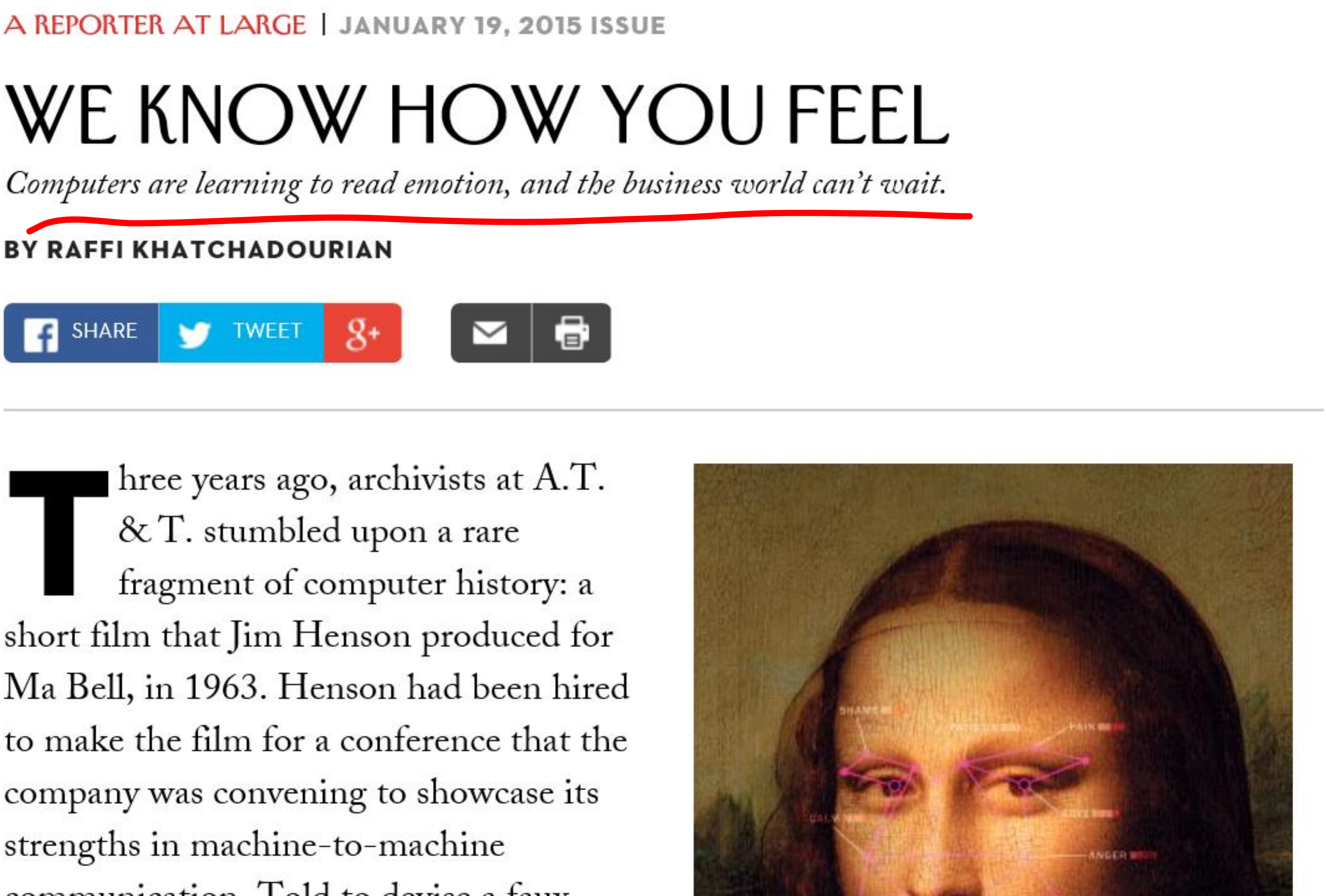
Towards Bridging the Gap

- Multicalibration
  - Approximate *calibration** for every large set defined by a circuit in a predetermined set C. Can post-process any predictor to obtain a multicalibrated one with no accuracy loss.
    - *Among those rated $v \in [0,1]$ by the predictor, the fraction who truly have a positive label is close to v; (Almost) correct *on average* for those rated v
  - Algorithm is a form of boosting for the circuits in C
    - Start with one predictor; find a “slice” S_v of a set $S \in C$ on which calibration is not satisfied; modify the predictor’s behavior on that slice to obtain a new predictor
    - These “slices” are defined adaptively!
      - Need more samples? No; instead use DP to identify miscalibrated slices

Hebert-Johnson, Kim, Reingold, Rothblum 2017

Affective Computing and Emotional Manipulation

Your Name Here



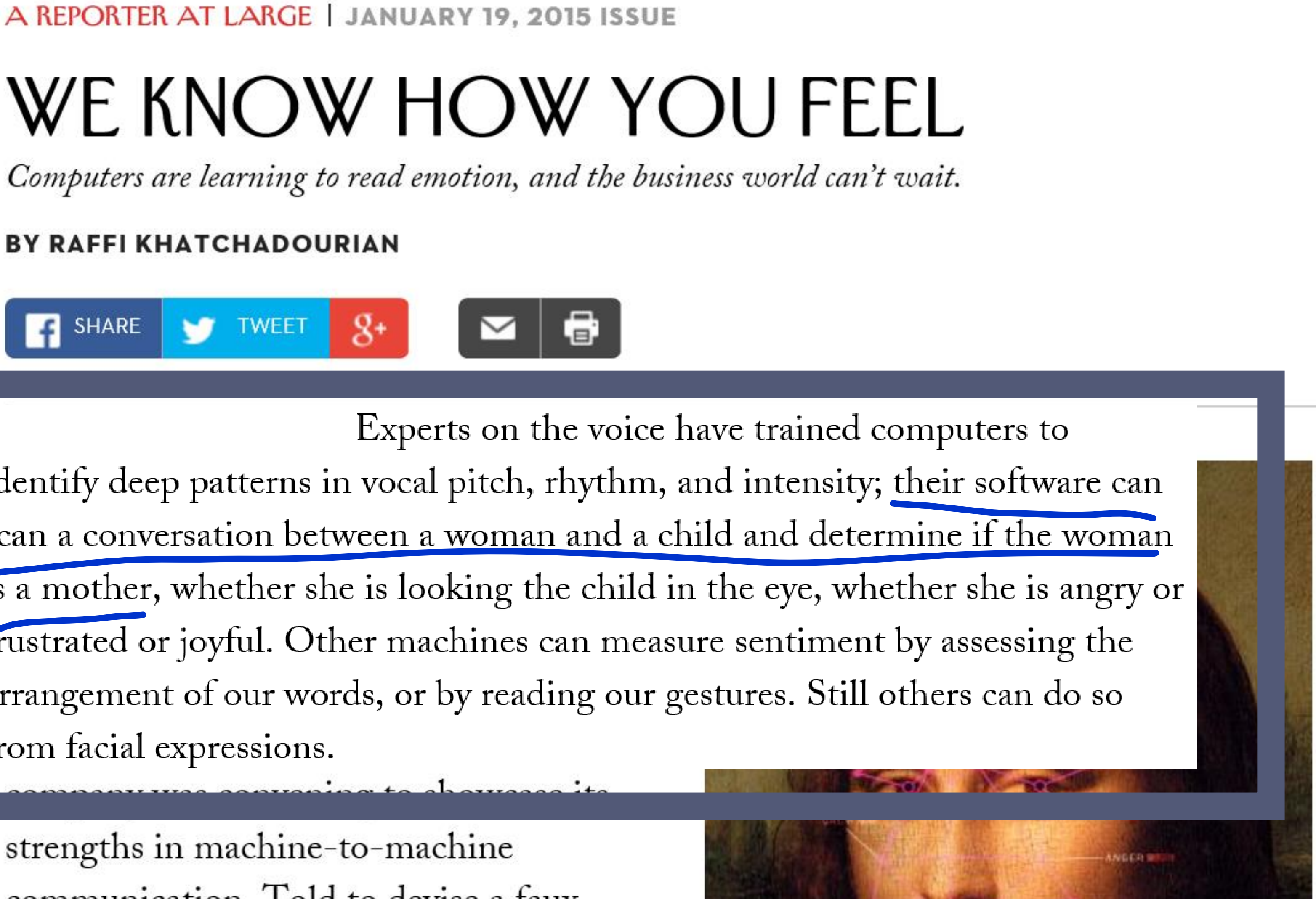
A REPORTER AT LARGE | JANUARY 19, 2015 ISSUE

WE KNOW HOW YOU FEEL

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

BY RAFFI KHATCHADOURIAN

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

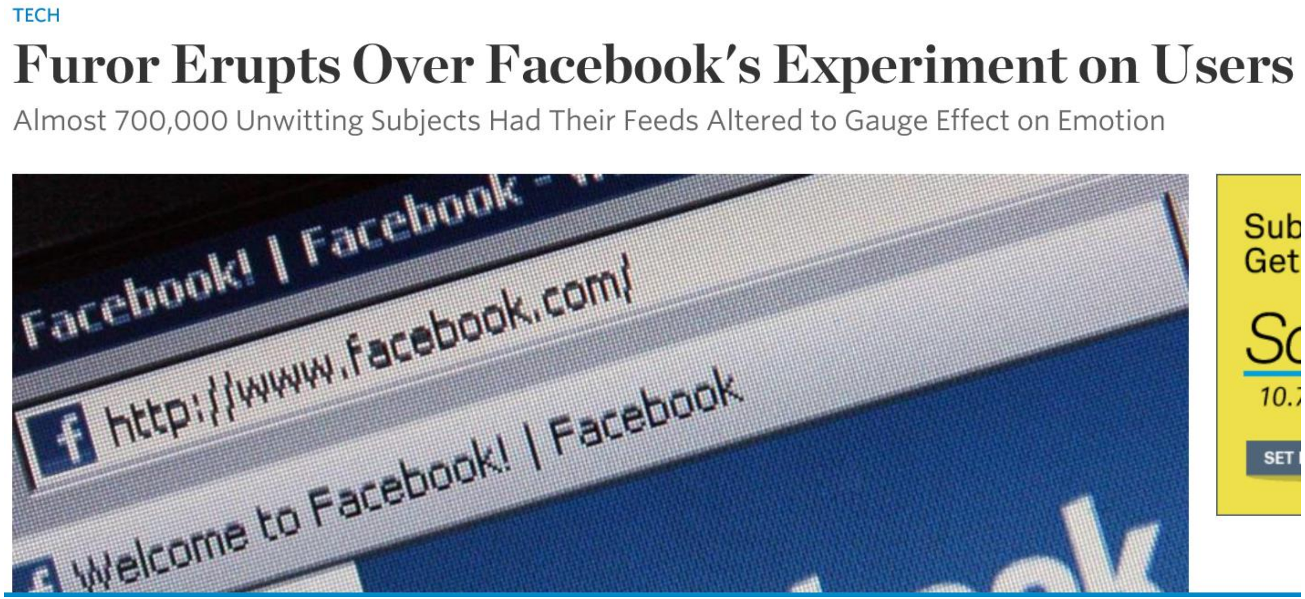
Computers are learning to read emotion, and the business world can't wait.

BY RAFFI KHATCHADOURIAN

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 or 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

Conclusions

- Work on the obvious questions
  - Make neural nets robust to adversarial examples
  - Develop the theory of differential privacy for social networks
  - Advance the state of the art in adaptive data analysis
  - Improve the state of the art for fairness under composition
  - Prevent the rewriting, by video manipulation, of history
  - NLP to make job ads appeal across genders

- Don’t *only* work on the obvious
  - Create a culture that *demands* signed video
  - Tackle fake news more broadly
  - Define and solve some aspect of the problem of manipulation via affective computing
  - Find a way to restore the informational commons; build a “dissonance engine”

Thank you!

Beyond CRYPTO, August 19, 2018