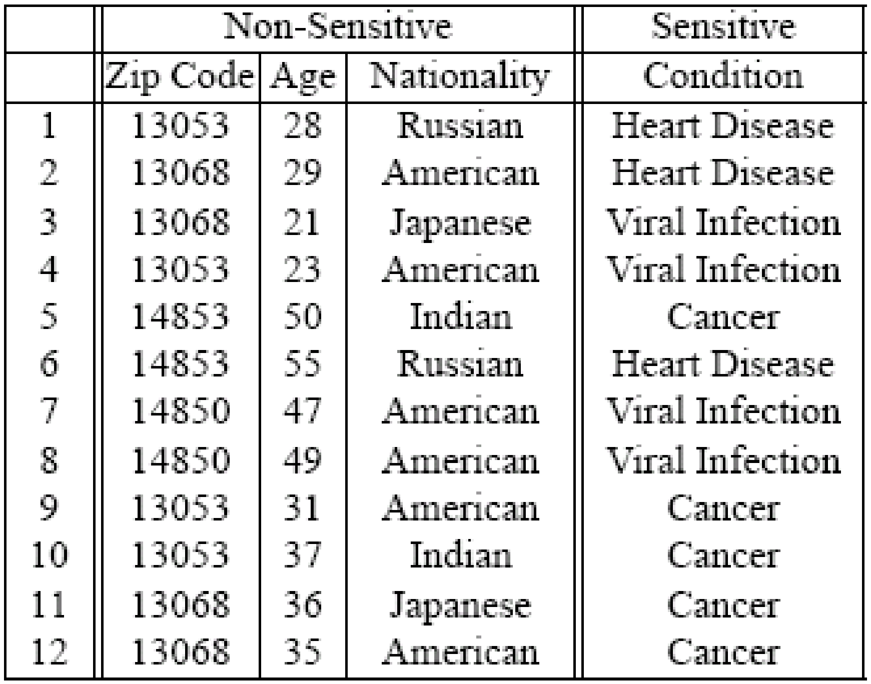
Administrative

- HW3 released
  - Due October 21
  - Question 2 requires installation of AdFisher
    - Start this question early!
- In-class discussion of privacy practices of organizations next Monday
  - Details on piazza

18734 Recitation

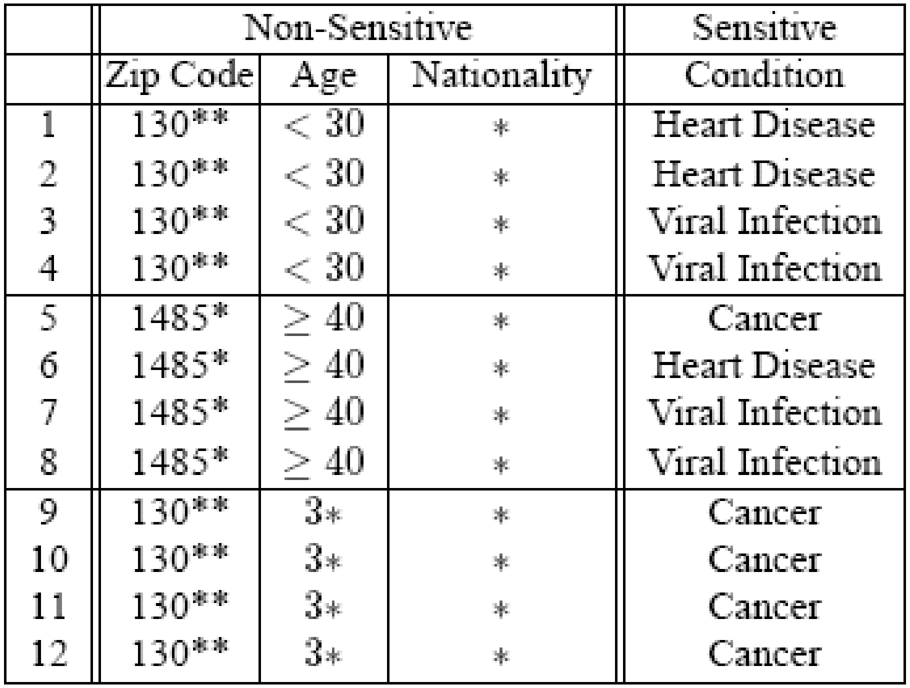
Distance Metrics

October 7, 2016



	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

Figure 1. Inpatient Microdata

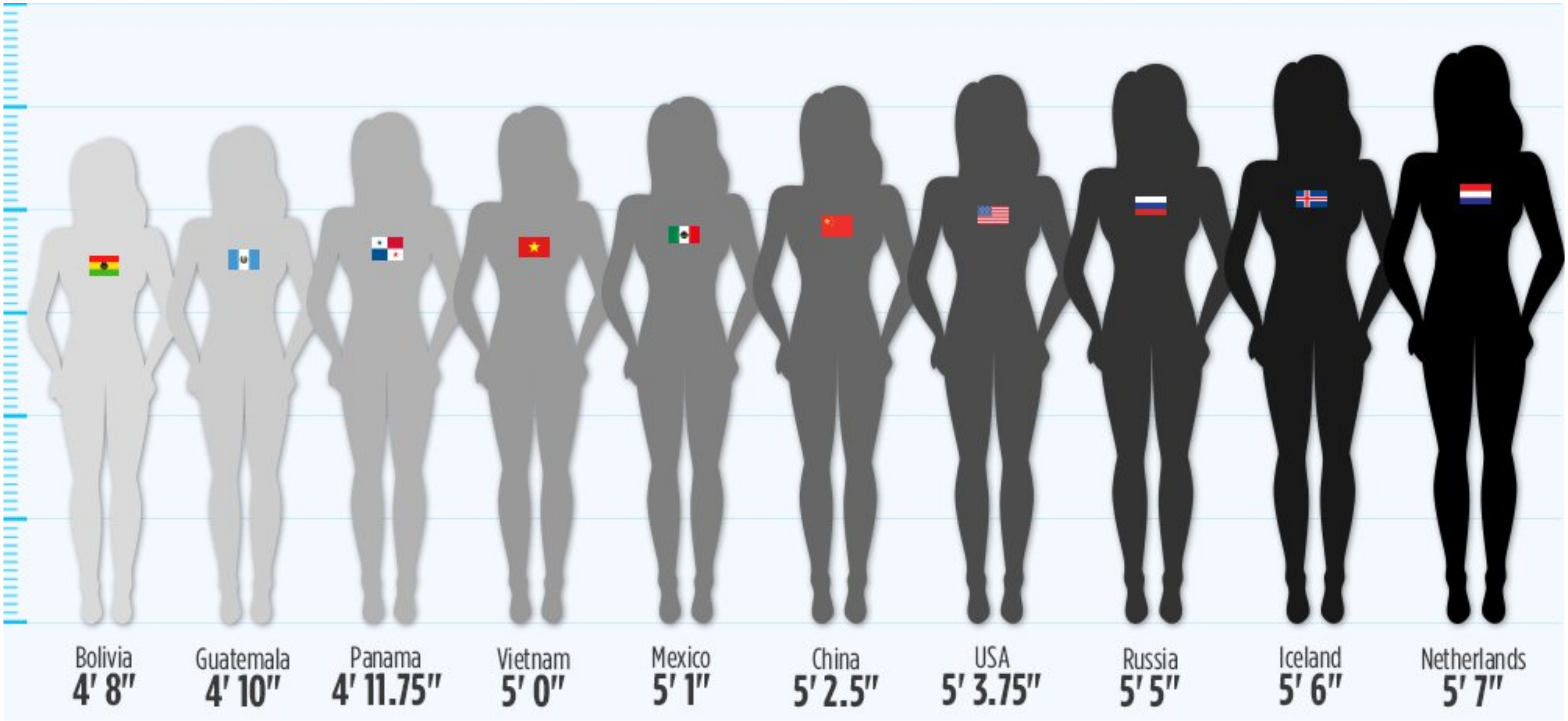


	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	$*$	Heart Disease
2	130**	< 30	$*$	Heart Disease
3	130**	< 30	$*$	Viral Infection
4	130**	< 30	$*$	Viral Infection
5	1485*	≥ 40	$*$	Cancer
6	1485*	≥ 40	$*$	Heart Disease
7	1485*	≥ 40	$*$	Viral Infection
8	1485*	≥ 40	$*$	Viral Infection
9	130**	$3*$	$*$	Cancer
10	130**	$3*$	$*$	Cancer
11	130**	$3*$	$*$	Cancer
12	130**	$3*$	$*$	Cancer

Figure 2. 4-anonymous Inpatient Microdata

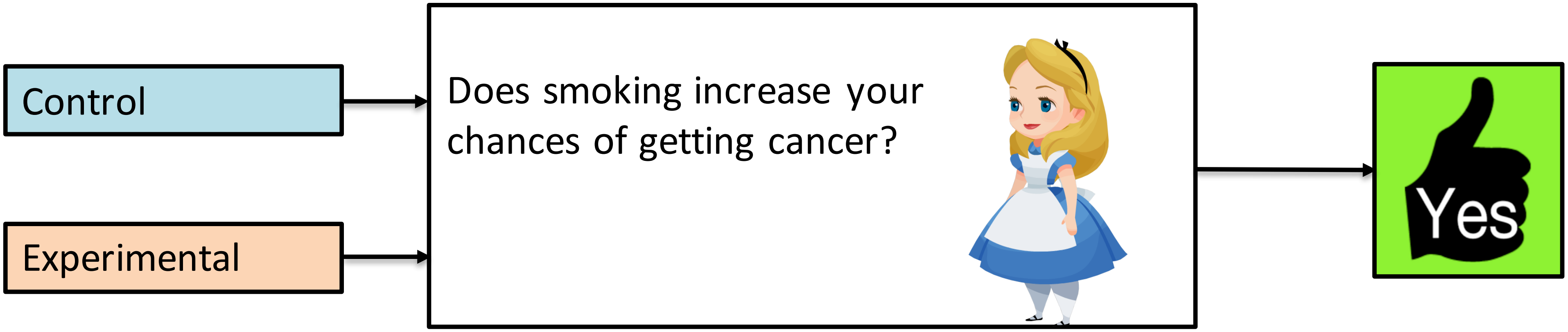
Goal of Statistical Disclosure Control

Reveal accurate statistics about a population while preserving the privacy of individuals



Stephanie Sun is one inch shorter than the average Russian woman

Differential Privacy

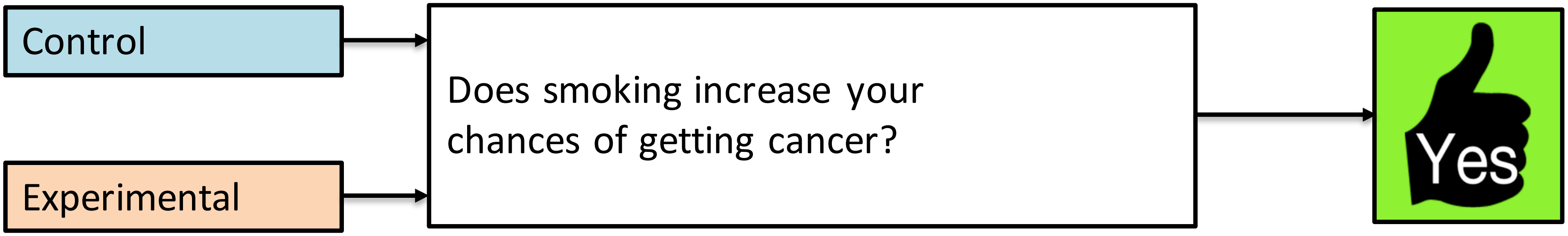


Control

Experimental

Does smoking increase your chances of getting cancer?

Yes

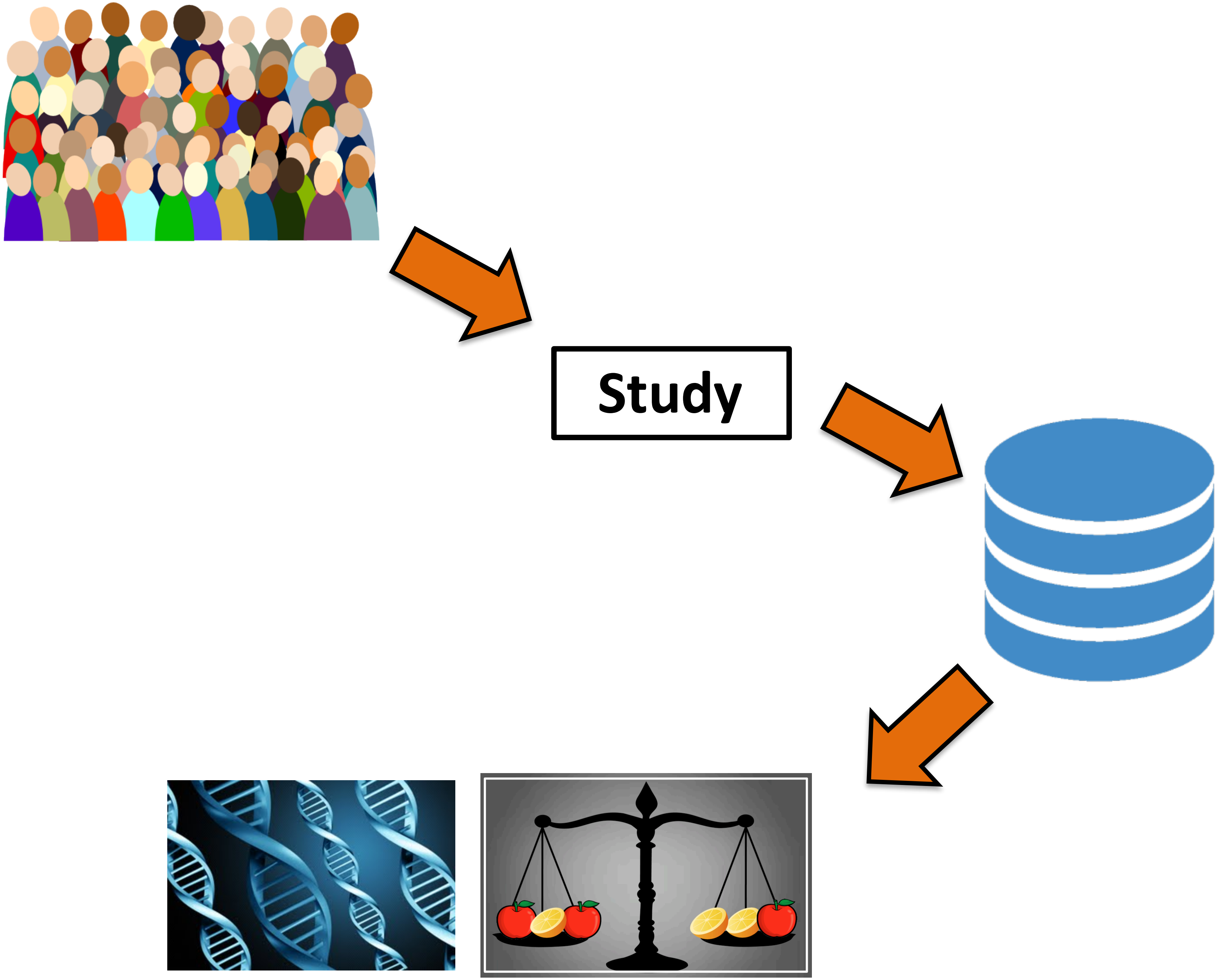


Control

Experimental

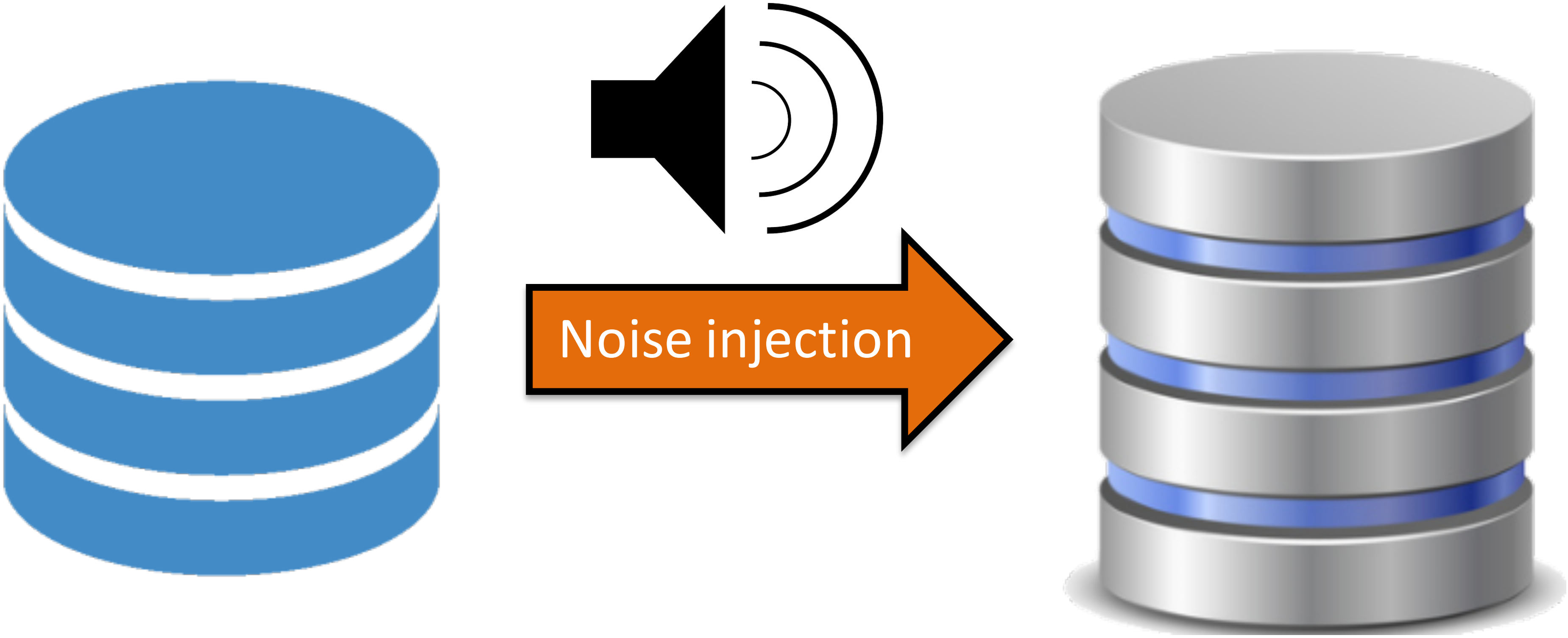
Does smoking increase your chances of getting cancer?

Yes



Study

Input perturbation



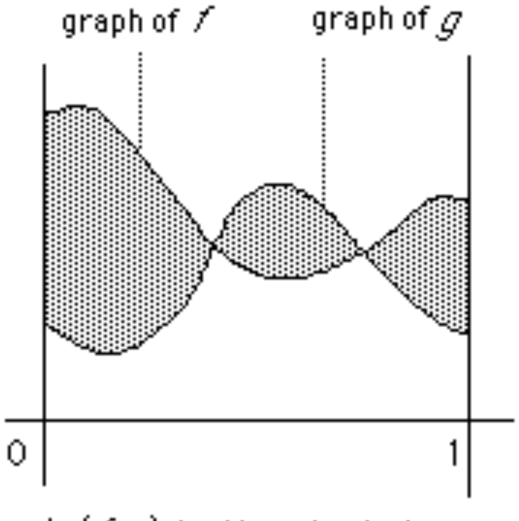
L1 Distance

- Between two points
  - $(x_1, x_2, \ldots, x_n)$ and $(y_1, y_2, \ldots, y_n)$
- $\sum_i |x_i - y_i|$

Distance between functions

- Between two discrete functions
  - $m_1(x)$, $m_2(x)$
  - $x \in \{x_1, x_2, \ldots, x_n\}$
  - $\sum_i |m_1(x_i) - m_2(x_i)|$

- Between two continuous functions
  - $n_1(y)$, $n_2(y)$
  - $y \in [y_1, y_2]$
  - ${}_{y1}\int^{y2} |n_1(y) - n_2(y)| dy$



$d_1(f, g)$ is the shaded area

Distance between probability distributions

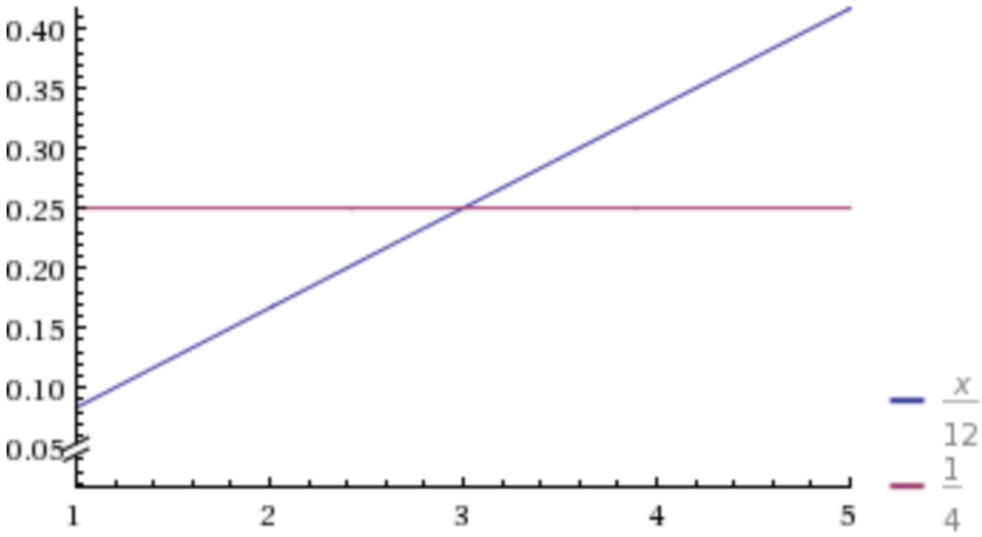
- Between two discrete distributions
  - PMFs $p_1(x)$, $p_2(x)$
  - $x \in \{x_1, x_2, \ldots, x_n\}$
  - $\sum_i |p_1(x_i) – p_2(x_i)|$

- Between two continuous distributions
  - PDFs $f_1(y)$, $f_2(y)$
  - $y \in [y_1, y_2]$
  - ${}_{y1}\int^{y2} |f_1(y) - f_2(y)|dy$

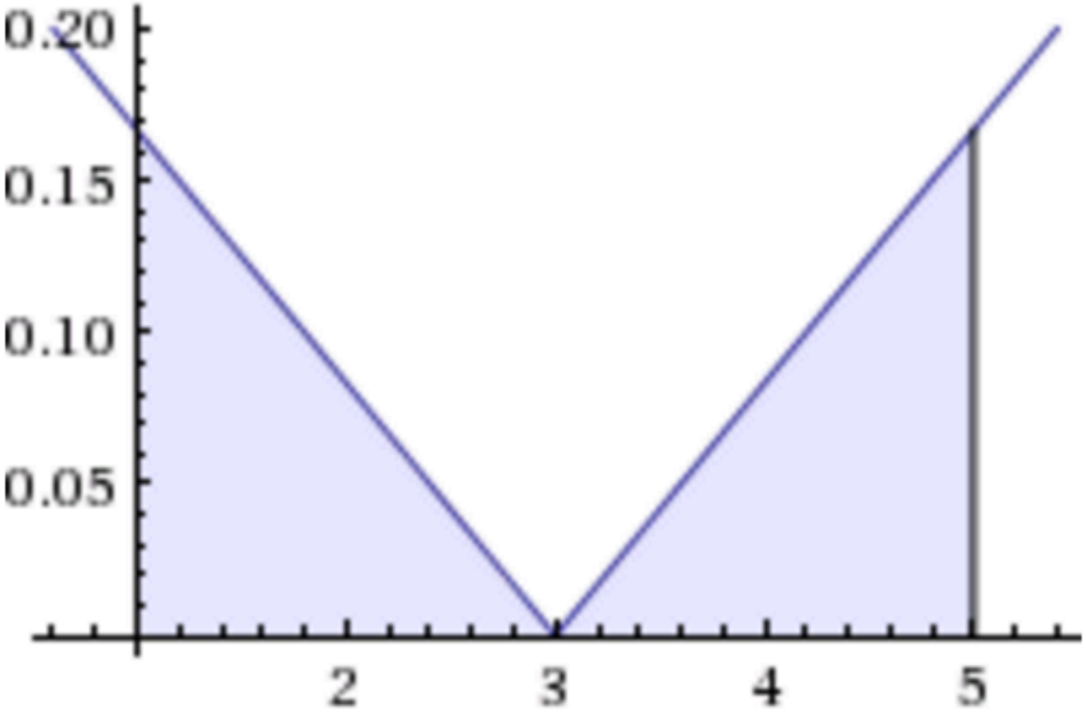
Exercise

- Find L1 distance between the following continuous distributions:
  - – $f_1(x) = x/12 \quad x \in [1, 5]$
  - – $f_2(x) = 1/4 \quad x \in [1, 5]$

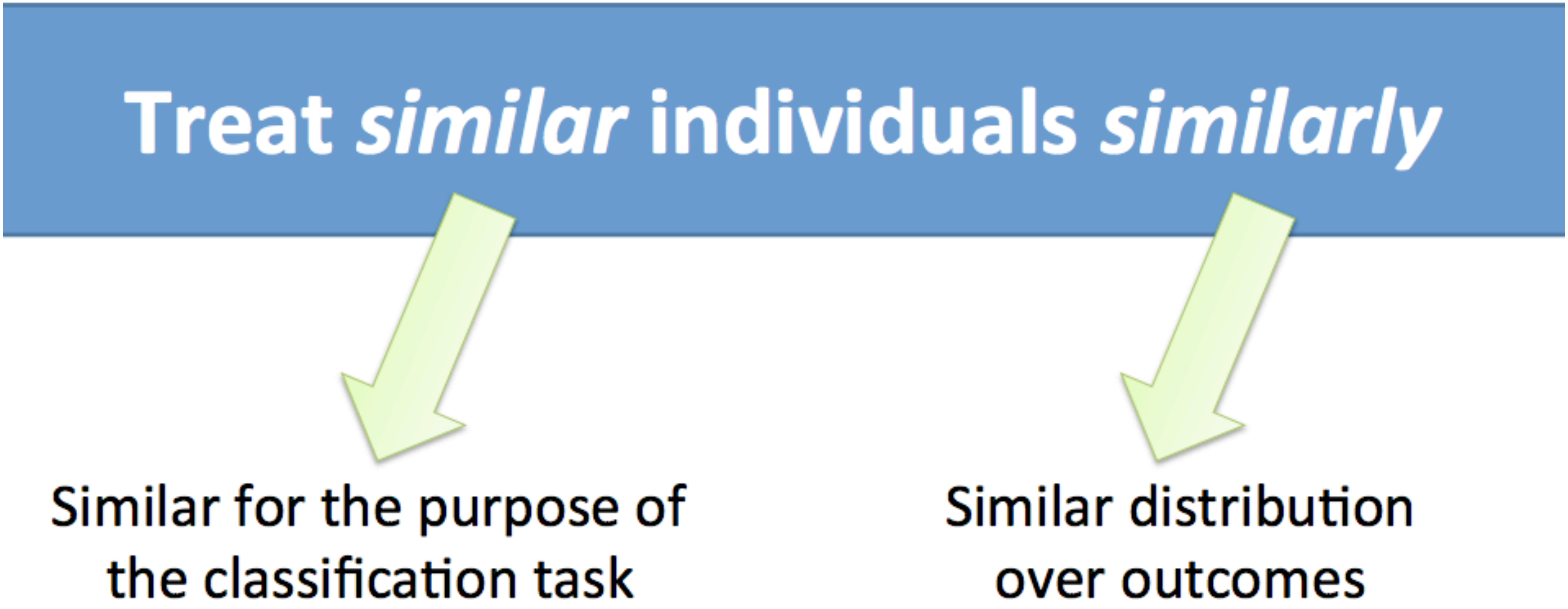
Plot:



Solution: 1/3



Individual Fairness



Treat *similar* individuals *similarly*

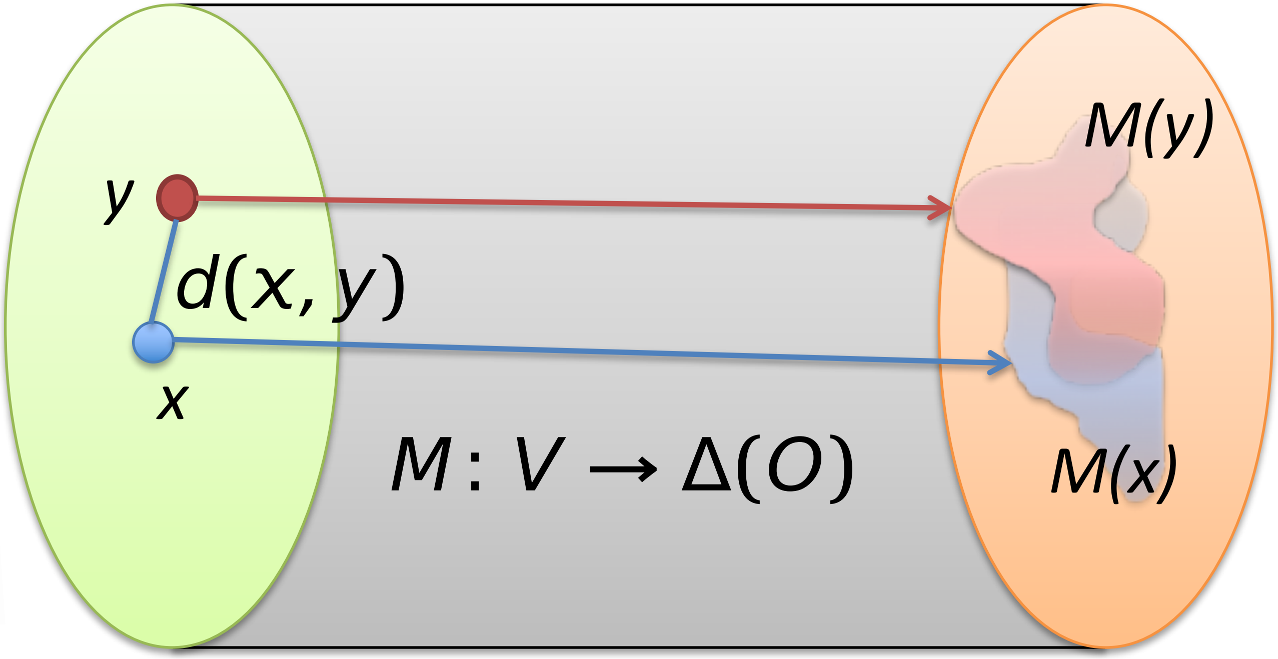
Similar for the purpose of the classification task

Similar distribution over outcomes

Fairness through Awareness

Metric $d : V \times V \to \mathbb{R}$

Lipschitz condition $\|M(x) - M(y)\| \leq d(x, y)$



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

V: Individuals

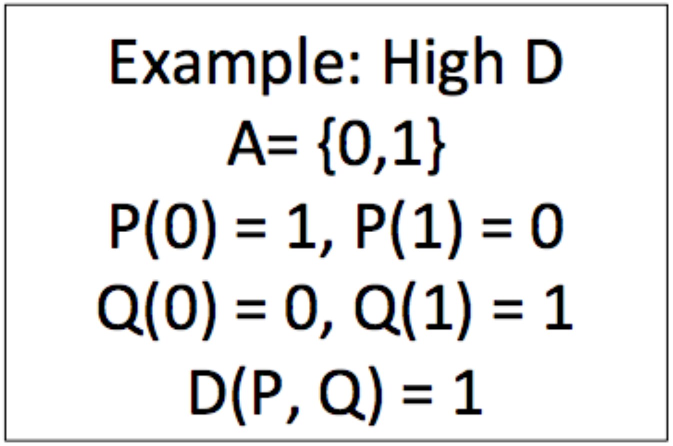
O: outcomes

Statistical Distance

- P, Q are probability measures on a finite domain A.
- Statistical distance between P and Q is:
  - $D(P,Q) = \frac{1}{2}\sum_{a\in A} |P(a) - Q(a)|$
    - where M(x)=P, M(y)=Q, O=A

Statistical Distance

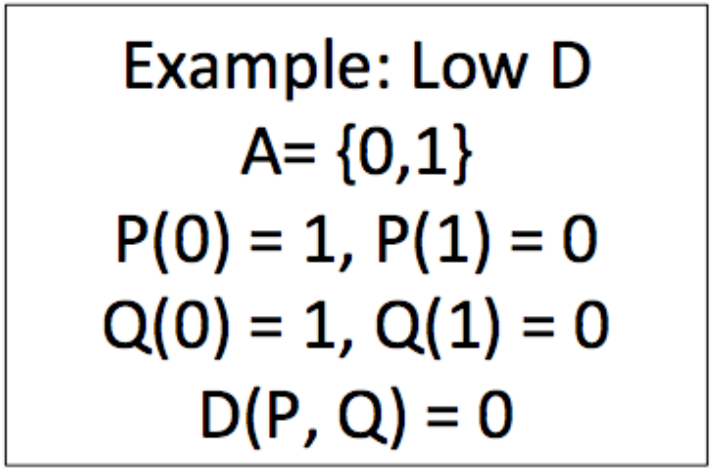
- P, Q are probability measures on a finite domain A.
- Statistical distance between P and Q is:
  - $D(P,Q) = \frac{1}{2}\sum_{a\in A} |P(a) - Q(a)|$
    - where M(x)=P, M(y)=Q, O=A



Example: High D
A= {0,1}
P(0) = 1, P(1) = 0
Q(0) = 0, Q(1) = 1
D(P, Q) = 1

Statistical Distance

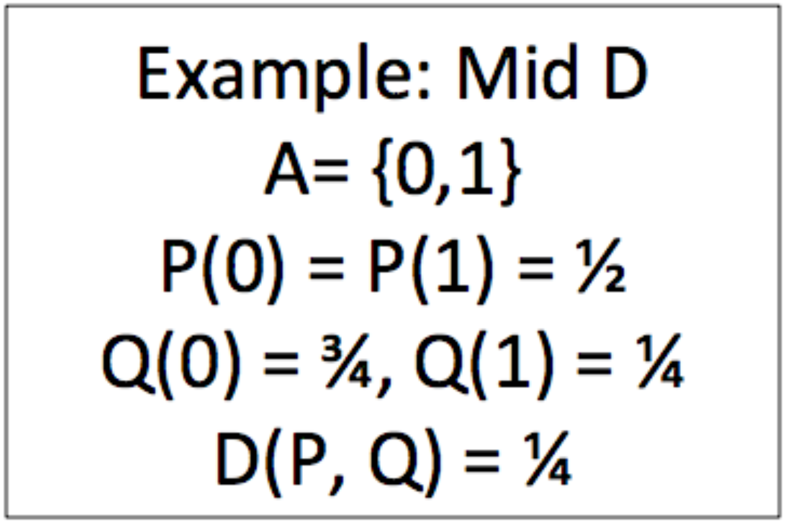
- P, Q are probability measures on a finite domain A.
- Statistical distance between P and Q is:
  - $D(P,Q) = \frac{1}{2}\sum_{a\in A}|P(a) - Q(a)|$
    - where M(x)=P, M(y)=Q, O=A



Example: Low D
A= {0,1}
P(0) = 1, P(1) = 0
Q(0) = 1, Q(1) = 0
D(P, Q) = 0

Statistical Distance

- P, Q are probability measures on a finite domain A.
- Statistical distance between P and Q is:
  - $D(P,Q) = \frac{1}{2}\sum_{a\in A}|P(a) - Q(a)|$
    - where M(x)=P, M(y)=Q, O=A



Example: Mid D

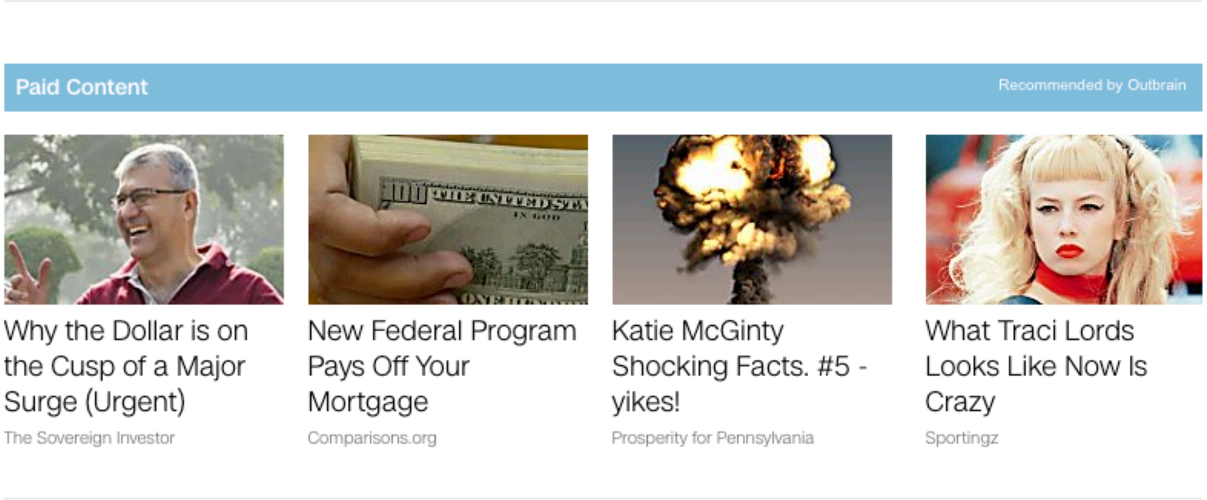
A= {0,1}

P(0) = P(1) = ½

Q(0) = ¾, Q(1) = ¼

D(P, Q) = ¼

Installing and Running AdFisher



Paid Content

Recommended by Outbrain

Why the Dollar is on the Cusp of a Major Surge (Urgent)
The Sovereign Investor

New Federal Program Pays Off Your Mortgage
Comparisons.org

Katie McGinty Shocking Facts. #5 - yikes!
Prosperity for Pennsylvania

What Traci Lords Looks Like Now Is Crazy
Sportingz

Setting up the environment

- AdFisher has been tested on Ubuntu 16.04 with Firefox 45.
- Use a VM if you are running Windows or Mac)
  - https://www.virtualbox.org/wiki/Downloads
- Ubuntu
  - https://www.ubuntu.com/download/desktop

Downgrade Firefox to Version 45

```
firefox --version Mozilla Firefox 47.0
apt-get remove firefox
wget
https://ftp.mozilla.org/pub/firefox/releases/45.0/linux-x86_64/en-US/firefox-45.0.tar.bz2
tar -xjf firefox-45.0.tar.bz2
mv firefox /opt/firefox45
ln -s /opt/firefox45/firefox /usr/bin/firefox firefox --version Mozilla Firefox 45.0
```

Reference: http://stackoverflow.com/questions/37761668/cant-open-browser-with-selenium-after-firefox-update

Installing the AdFisher

- Clone the git repository
  - https://github.com/tadatitam/info-flow-experiments
- Follow the instructions to install the python packages AdFisher uses:
  - https://github.com/tadatitam/info-flow-experiments/tree/master/AdFisher

Testing AdFisher

- Cd into AdFisher/examples
- Run **python demo_exp.py**