CS 500: Fundamentals of Databases

Data, Responsibly
Introduction to Data Mining

Julia Stoyanovich (*stoyanovich@drexel.edu*)

Data for and about people

facebook
Gmail
R
Dropbox
freshdirect
foursquare
amazon
GitHub

Google Calendar
twitter
kindle
CHASE
skype
NETFLIX
UnitedHealthcare
A UnitedHealth Group Company
iTunes
Google+
Network Attached Storage
Google Docs


StateFarm
Picasa
tripadvisor

The promise of big data

Power

Data collection capabilities

Big data: 5Vs (volume, velocity, variety, veracity, value)

enormous computational power

massively parallel processing

Opportunity

improve people's lives, e.g., recommendation

accelerate scientific discovery, e.g., medicine

boost innovation, e.g., autonomous cars

transform society, e.g., open government

optimize business, e.g., advertisement targeting

goal - progress

Illustration: big data and health

Analysis of a person's medical data, genome, social data

personalized medicine

personalized care and predictive measures

personalized insurance

expensive, or unaffordable, for those at risk

the same technology makes both possible!

Is data analysis objective or impartial?

Big data is algorithmic, therefore it cannot be biased! And yet…

- All traditional evils of **discrimination**, and many new ones, exhibit themselves in the big data eco system
- **Bias** that is inherent in the data or in the process, and that is often due to systemic discrimination, is propelled and amplified
- We need novel technological solutions to identify and rectify **irresponsible data analysis practices**
- Technology alone won't do: also need **policy**, **user involvement** and **education** efforts

http://www.allenovery.com/publications/en-gb/Pages/Protected-characteristics-and-the-perception-reality-gap.aspx

Data, responsibly

Because of its tremendous **power**, massive data analysis must be used **responsibly**

Fairness Diversity Transparency

we focus on fairness today, in a specific interpretation

Fairness is lack of bias

- What are the tasks we are interested in?
  - predictive analytics
- Where does bias come from?
  - data collection and analysis
- Analogy - scientific data analysis
  - collect a representative sample
  - do sound reproducible analysis
  - explain data collection and analysis
  - validate results

when data is about people, bias can lead to discrimination

The evils of discrimination

Disparate treatment is the illegal practice of treating an entity, such as a creditor or employer, differently based on a **protected characteristic** such as race, gender, age, religion, sexual orientation, or national origin.

Disparate impact is the result of systematic disparate treatment, where disproportionate **adverse impact** is observed on members of a **protected class**.

http://www.allenovery.com/publications/en-gb/Pages/Protected-characteristics-and-the-perception-reality-gap.aspx

Staples online pricing

THE WALL STREET JOURNAL.

WHAT THEY KNOW

Websites Vary Prices, Deals Based on Users' Information

By JENNIFER VALENTINO-DEVRIES, JEREMY SINGER-VINE and ASHKAN SOLTANI

December 24, 2012

It was the same Swingline stapler, on the same Staples.com website. But for Kim Wamble, the price was $15.79, while the price on Trude Frizzell's screen, just a few miles away, was $14.29.

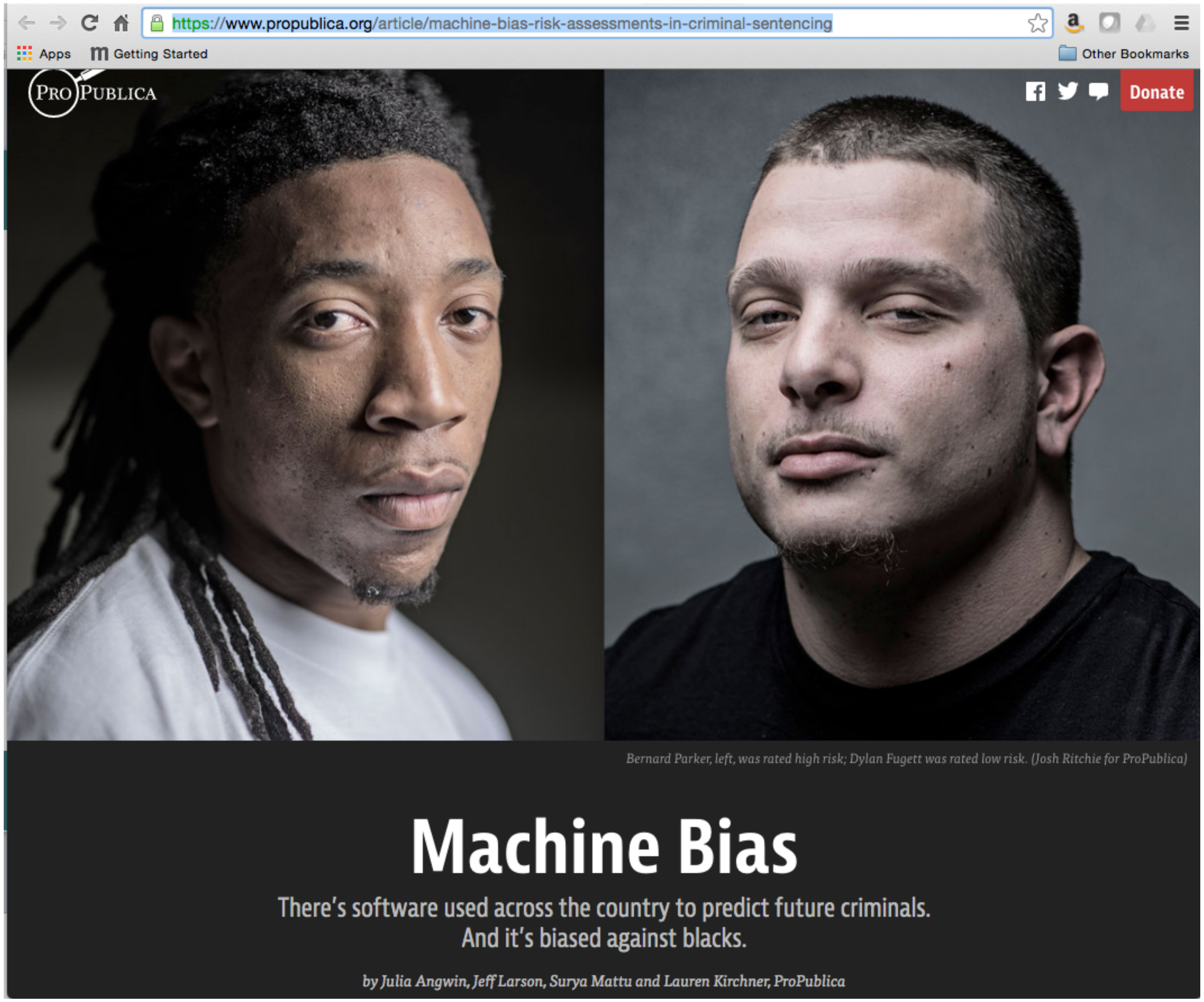
A key difference: where Staples seemed to think they were located.

WHAT PRICE WOULD YOU SEE?

lower prices offered to buyers who live in more affluent neighborhoods

data RESPONSIBLY

Racial bias in criminal sentencing



https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

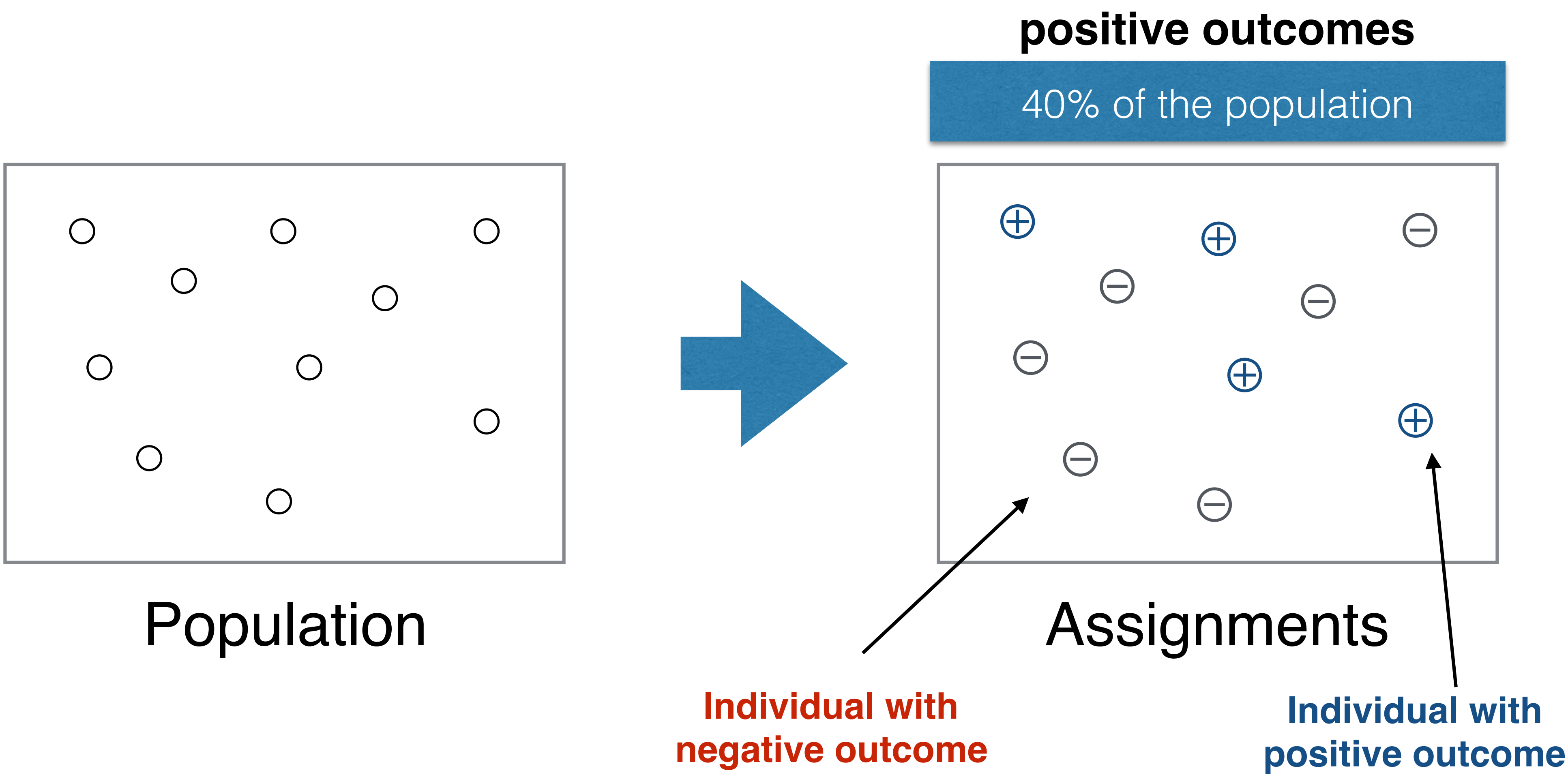
Outcomes

Consider a **vendor** assigning positive or negative **outcomes** to individuals.

Positive Outcomes	Negative Outcomes
offered employment	denied employment
accepted to school	rejected from school
offered a loan	denied a loan
offered a discount	not offered a discount

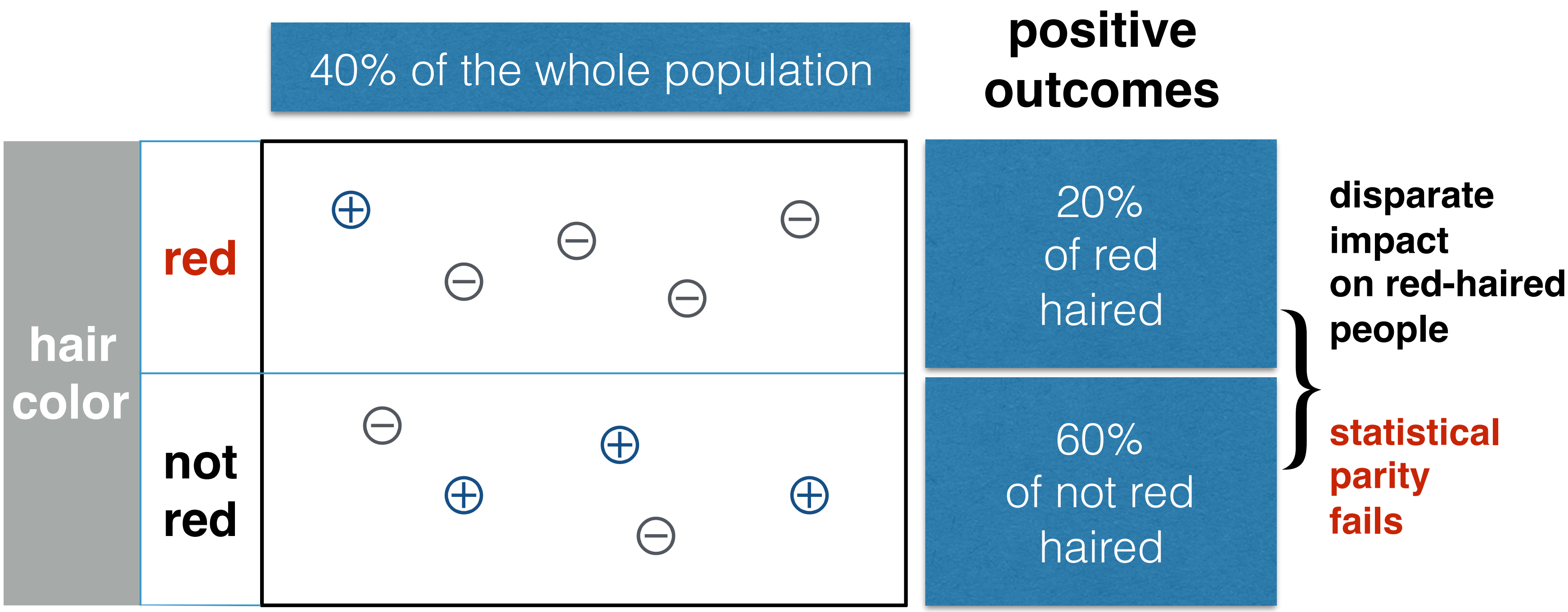
Assigning outcomes to populations

Fairness is concerned with how outcomes are assigned to a population



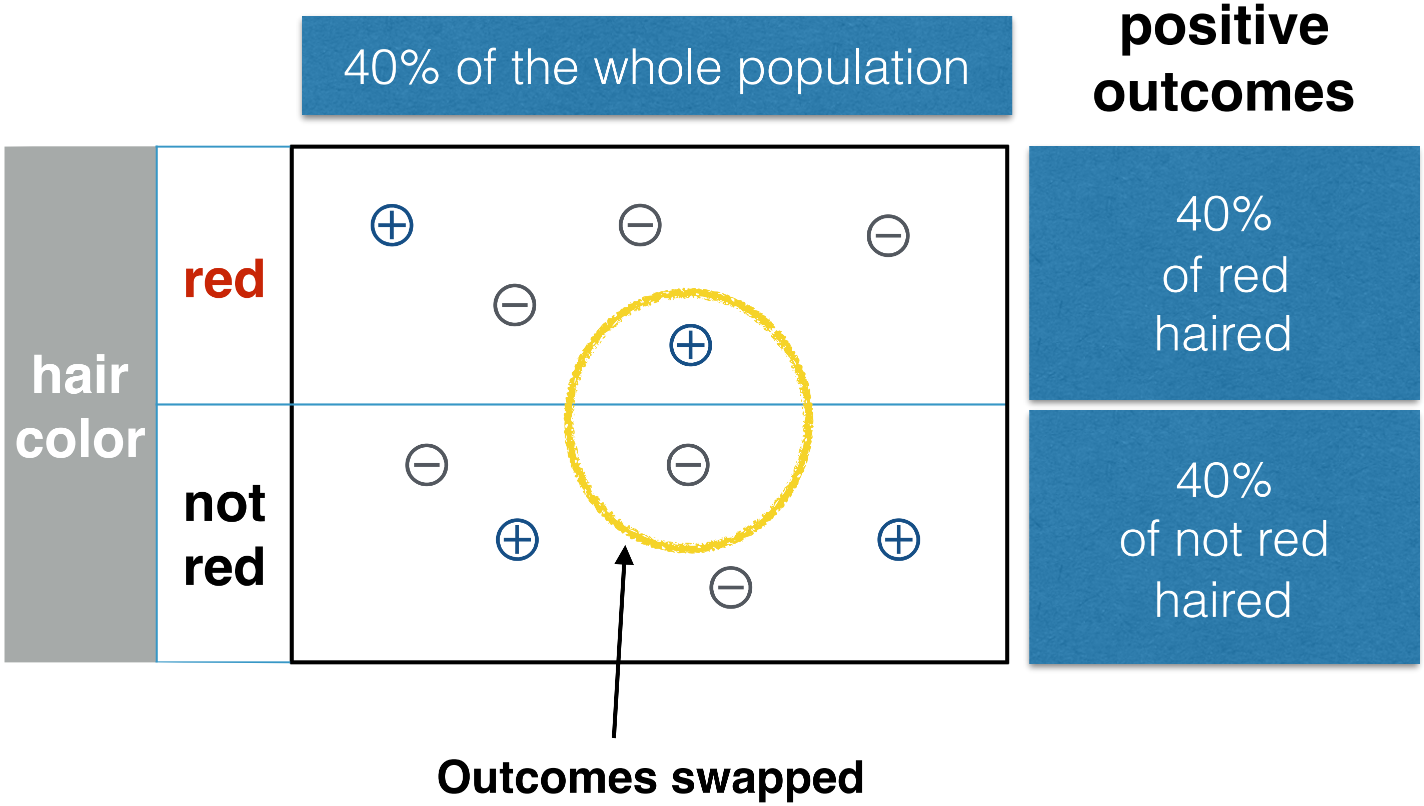
Sub-populations may be treated differently

Sub-population: those with red hair
(under the same assignment of outcomes)



Enforcing statistical parity

Statistical parity (aka **group fairness**)
demographics of the individuals receiving any outcome are the same as demographics of the underlying population



Redundant encoding

Now consider the assignments under both
hair color (protected) and **hair length** (innocuous)

hair color		hair length: long	hair length: not long
	red	⊕	⊖ ⊖ ⊖ ⊖
	not red	⊕ ⊕ ⊕	⊖ ⊖

positive outcomes

20% of red haired

60% of not red haired

Deniability
The vendor has adversely impacted red-haired people, but claims that outcomes are assigned according to hair length.

Blinding does not imply fairness

Removing **hair color** from the vendor's assignment process does not prevent discrimination!

		hair length	
		long	not long
hair color	red	⊕	⊖ ⊖ ⊖ ⊖
	not red	⊕ ⊕ ⊕	⊖ ⊖

positive outcomes

20% of red haired

60% of not red haired

Assessing disparate impact
Discrimination is assessed by the <u>effect</u> on the protected sub-population, not by the input or by the process that lead to the effect.

Redundant encoding

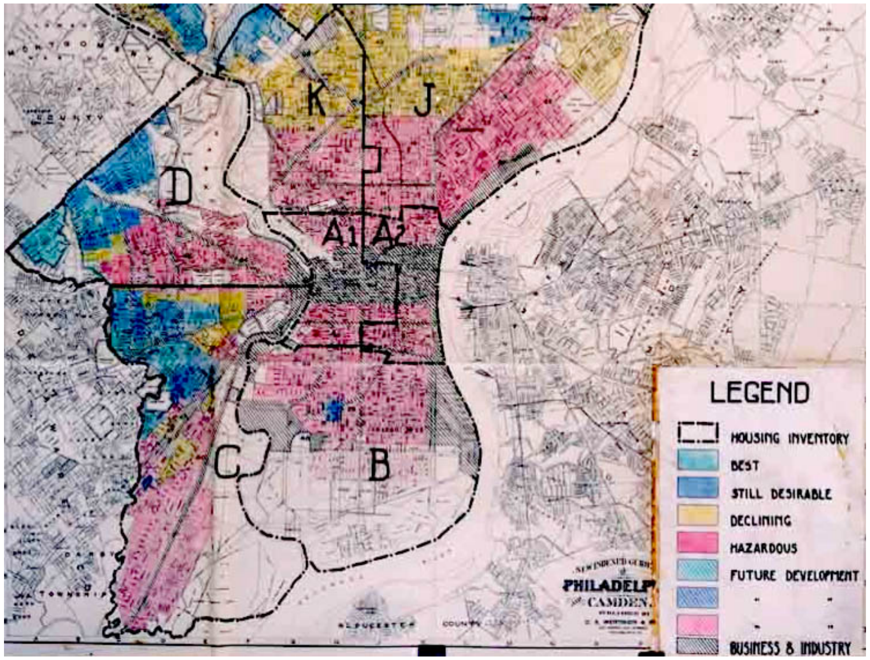
Let’s replace hair color with **race** (protected),
hair length with **zip code** (innocuous)

		zip code		**positive outcomes**
		10025	**10027**	
race	**black**	⊕	⊖ ⊖ ⊖ ⊖	20% of black
	white	⊕ ⊕ ⊕	⊖ ⊖	60% of white

The evils of discrimination

Redlining is the practice of arbitrarily denying or limiting financial services to specific neighborhoods, generally because its residents are people of color or are poor.

Philadelphia, 1936



wikipedia

Households and businesses in the red zones could not get mortgages or business loans.

Discrimination may be unintended

Staples website estimated user's location, **offering discounts** to those near rival stores, leading to discrimination w.r.t. to average income.

income		rival store proximity: close	rival store proximity: far
	low	⊕	⊖ ⊖ ⊖ ⊖
	high	⊕ ⊕ ⊕	⊖ ⊖

positive outcomes

20%
of low income

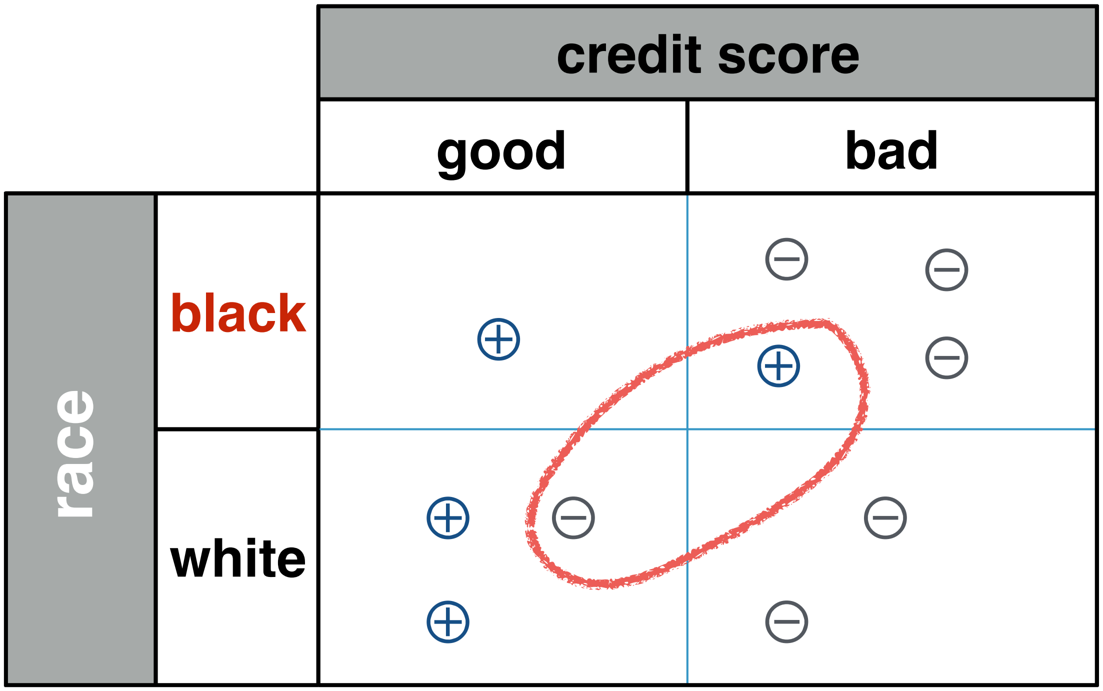
60%
of high income

Discrimination
Whether intentional or not, discrimination is unethical and, in many countries, illegal.

Imposing statistical parity

May be contrary to the goals of the vendor

positive outcome: offered a loan



positive outcomes

40% of black

40% of white

Impossible to predict loan payback accurately.
Use past information, which may itself be biased.

Defeating statistical parity

If the vendor wants to avoid offering positive outcomes to red-hairs, they can try to find a disqualifying secondary attribute.

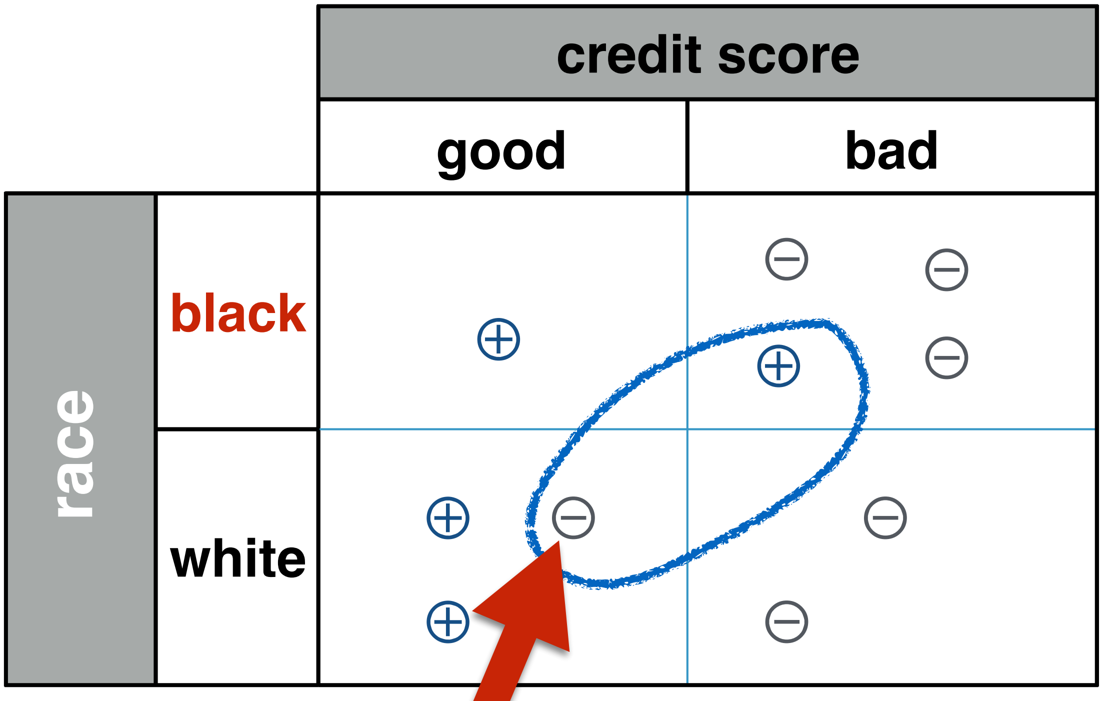
positive outcome: burger discount

		diet	
		vegetarian	**carnivore**
hair color	**red**	⊕ ⊕	⊖ ⊖ ⊖
	not red	⊖ ⊖	⊕ ⊕ ⊖

offered	**accepted**
40% of red haired	0% of red haired
40% of not red haired	40% of not red haired

Is statistical parity sufficient?

Statistical parity (aka **group fairness**)
demographics of the individuals receiving any outcome are the same as demographics of the underlying population



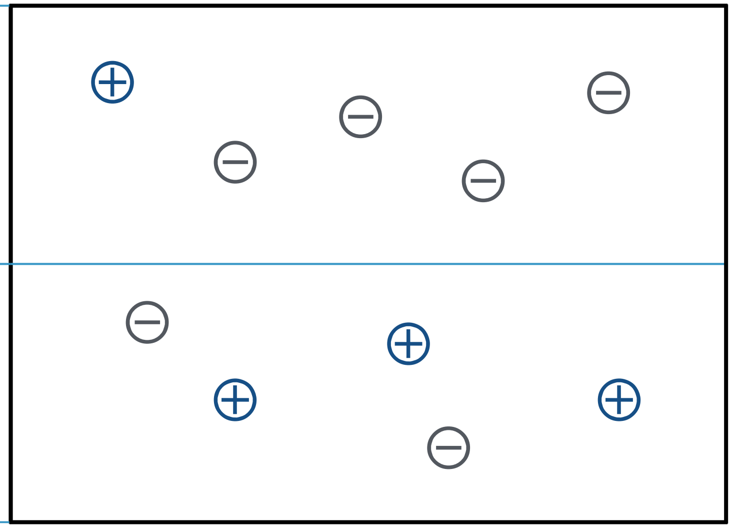
positive outcomes	**offered credit**
40% of black	
40% of white	

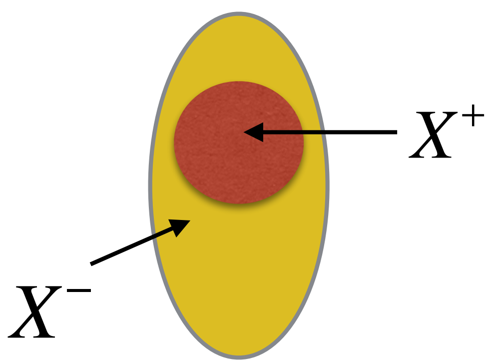
Individual fairness
any two individuals who are similar w.r.t. a particular task should receive similar outcomes

How do we quantify discrimination?

40% of the whole population

positive outcomes $Y = 1$

hair color		positive outcomes	
red		20% of red hair	$Y = 1 \mid X^+$
not red		60% of not red hair	$Y = 1 \mid X^-$



discrete (binary) protected feature *S*

X^+ are members of X with $S=1$

X^- are members of X with $S=0$

Let’s make things concrete

Introduction to data mining

supplementary material:
http://infolab.stanford.edu/~ullman/mmds/ch1.pdf
http://infolab.stanford.edu/~ullman/mmds/ch6.pdf

Big data according to T.S. Eliot

Choruses from *The Rock* (1934)

The Eagle soars in the summit of Heaven,
The Hunter with his dogs pursues his circuit.
O perpetual revolution of configured stars,
O perpetual recurrence of determined seasons,
O world of spring and autumn, birth and dying!

biology

astronomy

climate & weather

population dynamics

1888-1965

The endless cycle of idea and action,
Endless invention, endless experiment,
Brings knowledge of motion, but not of stillness;
Knowledge of speech, but not of silence;
Knowledge of words, and ignorance of the Word.
All our knowledge brings us nearer to death,
But nearness to death no nearer to God.

scientific experimentation

Where is the Life we have lost in living?
Where is the wisdom we have lost in knowledge?
Where is the knowledge we have lost in information?

The cycles of Heaven in twenty centuries
Brings us farther from God and nearer to the Dust.

More optimism for the 21st century!

Knowledge discovery and data mining

"Knowledge discovery in databases is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data."

"Data mining is a step in the KDD process consisting of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns over the data."

Fayyad et al., 1996.

Knowledge discovery and data mining

- Why we need data mining
  - "Drowning in data yet starving for knowledge", anonymous
  - "Computers have promised us a fountain of wisdom but delivered a flood of data", W. J. Frawley, G.Piatetsky-Shapiro, and C. J. Matheus
  - “Where is the wisdom we have lost in knowledge? Where is the knowledge we have lost in information?”, T. S. Eliot
- What data mining is not
  - Data mining, noun: "Torturing data until it confesses ... and if you torture it enough, it will confess to anything", Jeff Jonas, IBM
  - "An unethical econometric practice of massaging and manipulating the data to obtain the desired results", W.S. Brown “Introducing Econometrics”

From http://www.cs.ccsu.edu/~markov/ccsu_courses/DataMining-1.html

Some types of data mining

- Association rule mining - today's lecture
  - e.g., 72% of customers who bought cookies also bought milk

- Classification - related to association rule mining, today's lecture
  - e.g., is a new customer applying for a loan a good investment or not?

    If STATUS = married and INCOME > 50K and HOUSE_OWNER = yes

    Then GRANT_LOAN = yes

- Finding sequential / temporal patterns
  - e.g., find the set of genes that are differentially expressed, and whose expression precedes the onset of a disease

- Clustering
  - Similar to classification, but classes are not known ahead of time

Association rule mining

- Proposed by Agrawal, Imielinski and Swami in SIGMOD 1993
- The now-classic Apriori algorithm by Agrawal and Srikant was published in VLDB 1994, received the 10-year best paper award at VLDB 2004
- Initially used for market basket data analysis, but has many other applications
- Answers two related questions
  1. Which items are often purchased together?
     - frequent itemsets, e.g., Milk, Cookies
     - have an associated **support**
  2. Which items will likely be purchased, based on other purchased items?
     - association rules, e.g., Diapers => Beer
     - meaning: if diapers are bought in a transaction, beer is also likely bought in the same transaction.
     - each association rule is derived from two frequent itemsets
     - have an associated **support** and **confidence**

The model: market-basket data

- $I = \{i_1, i_2, \ldots, i_m\}$ is the set of available items, e.g., a product catalog of a store
- Transaction t is a set of items purchased together, $t \subseteq I$, has a transaction id (TID)

  t_1: {bread, cheese, milk}

  t_2: {apple, eggs, salt, yogurt}

  t_3: {biscuit, cheese, eggs, milk}

- Database T is a set of transactions $\{t_1, t_2, \ldots, t_n\}$

What is not represented by this model?

Itemsets

$X \subset I$ is an **itemset**

- $X = \{milk, bread, cereal\}$ is an itemset
- X is a 3-itemset (a k-itemset with k=3)
- X has **support** supp if supp% of transactions contain X

A transaction t contains an itemset X if $X \subseteq t$

- t is said to give support to X

A user specifies a support threshold minSupp

Itemsets with support $\geq$ minSupp are **frequent itemsets**

Example

TID	Items
1	A
2	A C
3	A B D
4	A C
5	A B C
6	A B C

minSupp = 20% at least 2 transactions

How many possible item sets are there?
$2^4 = 16$

	itemset	support
★	A	100%
★	B	50%
★	C	67%
	D	17%
★	A B	50%
★	A C	67%
	A D	17%
★	B C	33%
	B D	17%
	C D	0
★	A B C	33%
	A B D	17%
	B C D	0
	A C D	0
	A B C D	0

Association Rules

An **association rule** is an implication $X \rightarrow Y$, where X, $Y \subset I$, *and* $X \cap Y = \emptyset$

example: {milk, bread}$\rightarrow${cereal}

"A customer who purchased X is also likely to have purchased Y in the same transaction"

we are interested in rules with a **single item** in Y

can we represent {milk, bread} -> {cereal, cheese}?

Rule $X \rightarrow Y$ holds with **support** *supp* in T if *supp* % of transactions contain $X \cup Y$

$supp \approx \Pr(X \cup Y)$

Rule $X \rightarrow Y$ holds with **confidence** *conf* in T if *conf* % of transactions that contain X also contain Y

$conf \approx \Pr(Y \mid X)$

$conf\,(X \rightarrow Y) = supp\,(X \cup Y) \,/\, supp\,(X)$

Example

minSupp = 20%   at least 2 transactions

minConf = 75%

supp = 50%

$A \rightarrow B$ conf = 50% / 100% = 50%

$B \rightarrow A$ conf = 50% / 50% = 100%

supp = 50%

$B \rightarrow C$ conf = 100%

$C \rightarrow B$ conf = 75%

supp = 67%

$A \rightarrow C$ conf = 67%

$C \rightarrow A$ conf = 100%

supp = 33%

$AB \rightarrow C$ conf = 67%

$AC \rightarrow B$ conf = 50%

$BC \rightarrow A$ conf = 100%

itemset	support
A	100%
B	50%
C	67%
D	17%
A B	50%
A C	67%
A D	17%
B C	33%
B D	17%
C D	0
A B C	33%
A B D	17%
B C D	0
A C D	0
A B C D	0

Association Rule Mining

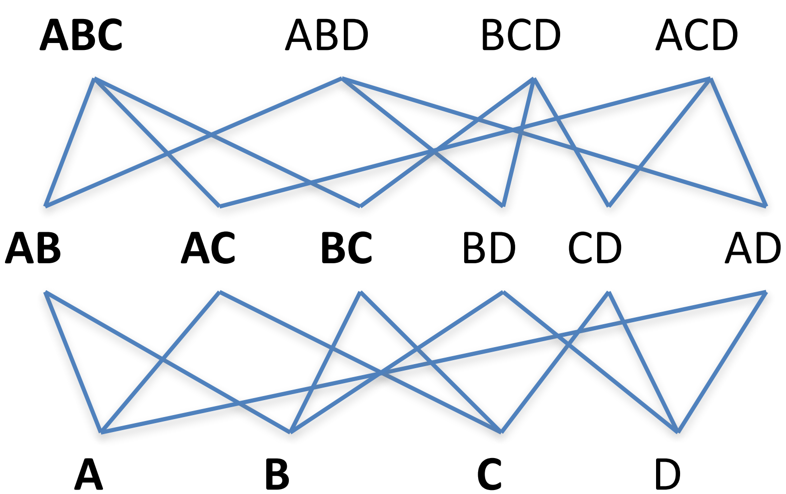
- Goal: find all association rules that satisfy the user-specified minimum support and minimum confidence

- Algorithm outline
  - Step 1: find all frequent itemsets
  - Step 2: find association rules

- Take 1: naïve algorithm for frequent itemset mining
  - Enumerate all subsets of I, check their support in T
  - What is the complexity?
  - Any obvious optimizations?

Downward Closure

- Recall: a frequent itemset has support $\geq$ minSupp
- Key idea: Use the downward closure property
  - — all subsets of a frequent itemset are themselves frequent
  - — conversely: if an itemset contains any infrequent itemsets as subsets, it cannot be frequent (we know this apriori)
  - — Is an itemset necessarily frequent if all its subsets are frequent?
    - No! supp(X U Y) $\leq$ min(supp(X), supp(Y))



	itemset	support
★	A	100%
★	B	50%
★	C	67%
	D	17%
★	A B	50%
★	A C	67%
	A D	17%
★	B C	33%
	B D	17%
	C D	0
★	A B C	33%
	A B D	17%
	B C D	0
	A C D	0
	A B C D	0

The *Apriori* Algorithm

Algorithm Apriori(T)

F_1 = {*frequent 1-itemsets*};

for ($k = 2$; $F_{k-1} \neq \varnothing$; k++) **do**

$C_k \leftarrow$ candidate-gen(F_{k-1});

for each transaction $t \in T$ **do**

for each candidate $c \in C_k$ **do**

if c is contained in t **then**

c.count++;

end

end

$F_k \leftarrow \{c \in C_k \mid c.count/n \geq minsup\}$

end

return $F \leftarrow \bigcup_k F_k$;

Apriori candidate generation

The candidate-gen function takes F_{k-1} and returns a superset (called the candidates) of the set of all frequent k-itemsets. It has two steps:

- **Join:** generate all possible candidate itemsets C_k of length k
- **Prune:** remove those candidates in C_k that have infrequent subsets
  - Which subsets do we check?

Apriori candidate generation

Assume a lexicographic ordering of the items

Join

Insert into C_k

Select $p.item_1, p.item_2, \ldots, p.item_{k-1}, q.item_{k-1}$

From F_{k-1} p, F_{k-1} q

Where $p.item_1 = q.item_1$

And $p.item_2 = q.item_2$

And ….

And $p.item_{k-1} < q.item_{k-1}$ Why not $p.item_{k-1} \neq q.item_{k-1}$?

Prune

for each c in C_k do

for each (k-1) subset s of c do

if (s not in F_{k-1}) then

delete c from C_k

Generating association rules

for each frequent *k*-itemset *X*

for each 1-itemset A ⊂ X

compute conf (X-A → A) = supp (X) / supp (X-A)

if conf(X-A → A) ≥ minConf then X-A → A is an association rule

see slide 34 for an example

How are association rules different from functional dependencies in relational databases?

Performance of *Apriori*

- The possible number of frequent itemsets is exponential, $O(2^m)$, where m is the number of items
- Apriori exploits sparseness and locality of data
  - Still, it may produce a large number of rules: thousands, tens of thousands, ….
  - So, thresholds should be set carefully. What are some good heuristics?
- Let's take another look at the algorithm

The *Apriori* Algorithm

```
Algorithm Apriori(T)
    F1 = {frequent 1-itemsets};
    for (k = 2; Fk-1 ≠ ∅; k++) do
        Ck ← candidate-gen(Fk-1);
        for each transaction t ∈ T do      // a full scan of T for each k!
            for each candidate c ∈ Ck do
                if c is contained in t then
                    c.count++;
            end
        end
        Fk ← {c ∈ Ck | c.count/n ≥ minsup}
    end
return F ← ⋃k Fk;
```

The *AprioriTid* Algorithm

Algorithm AprioriTid(T)

F_1 = {*frequent 1-itemsets*}; $T_1 = T$;

for ($k = 2$; $F_{k-1} \neq \emptyset$; k++) **do**

$C_k \leftarrow$ candidate-gen(F_{k-1});

$T_k = \{\}$

for each transaction $t \in T_{k-1}$ **do**

C_k^t = {*itemsets in* C_k *to which* t *gives support*}

for each candidate $c \in C_k^t$ **do**

c.count++;

end

$T_k = T_k \cup \langle t.TID, C_k^t \rangle$

end

$F_k \leftarrow \{c \in C_k \mid c.count/n \geq minsup\}$

end

return $F \leftarrow \bigcup_k F_k$;

AprioriTid example

minSupp = 30%

T1 = T

TID	set of itemsets
1	{{A}}
2	{{A}, {C}}
3	{{A}, {B}, {D}}
4	{{A}, {C}}
5	{{A}, {B}, {C}}
6	{{A}, {B}, {C}}

F1 = {A} {B} {C}

C2 = {AB} {AC} {BC}

T2

TID	set of itemsets
2	{{A C}}
3	{{A B}}
4	{{A C}}
5	{{AB}, {AC}, {BC}}
6	{{AB}, {AC}, {BC}}

F3 = {AB} {AC} {BC}

C3 = {ABC}

T3

TID	set of itemsets
5	{{A B C}}
6	{{A B C}}

Apriori vs. *AprioriTid*

Any guesses as to the relative performance?

The goal is to avoid scanning the database T

So, we are computing and carrying around a redundant data structure that contains a sub-set of T, in conveniently pre-processed form

When does this NOT help performance? For small k? For large k?

So, why the 10-year best paper award?

- Why is this such a big deal?
  - A fairly simple model
  - A fairly simple bottom-up algorithm
  - A fairly obvious performance optimization
  - No pretty optimality proof

- But this is only simple in hindsight! Plus….
  - The algorithm works well in practice
  - Many real applications
  - Many possible useful extensions

From association rules to classification

T: database of transactions (market-basket data)

TID	items
100	shirt
200	jacket, shoes, boots
300	pants, boots
400	shoes
500	shoes
600	jacket

D: database of profiles of individuals (dating, employment, credit, criminal)

UID	gender	age	score
Ann	F	31	low
Bob	M	27	high
Cate	F	55	high
Dave	M	43	low

TD: database of profiles of individuals, transformed to look like transactions

UID	attributes
Ann	gender=F, age ε [30,35), score=low
Bob	gender=M, age ε [25,30), score=high
Cate	gender=F, age ε [55, 60), score=high
Dave	gender=M, age ε [40, 45), score=low

Classification association rules (CARs)

D: database of individuals

UID	sex	age	score
Ann	F	31	**low**
Bob	M	27	**high**
Cate	F	55	**high**
Dave	M	43	**low**

TD: database of individuals that looks like transactions

UID	attributes
Ann	gender=F, age ε [30,35), **score=low**
Bob	gender=M, age ε [25,30), **score=high**
Cate	gender=F, age ε [55, 60), **score=high**
Dave	gender=M, age ε [40, 45), **score=low**

$S\ X \rightarrow C$ *X* is a set of attribute-value pairs, and *c* ε *C* is a (binary) outcome

in our example, *score* is the outcome (low or high), also called the class label

continuous attribute values must be discretized (mapped to buckets) as part of the transformation - age in our example

Potentially discriminatory rules (PD-CARs)

D: database of individuals

UID	gender (S)	age (X1)	edu (X2)	score (C)
Ann	**F**	[30,35)	BS	**low**
Bob	**M**	[25,30)	MS	**high**
Cate	**F**	[55, 60)	PhD	**high**
Dave	**M**	[40, 45)	BS	**low**

$S\ X \rightarrow C$

S is a binary attribute-value assignment -
membership in a protected group (gender in our example)

X is a set of "regular" attribute-value pairs
(age and edu in our example)

C is a binary attribute-value assignment -
classification outcome (score in our example)

Potentially discriminatory rules (PD-CARs)

R: $S\ X \rightarrow C$

UID	gender (S)	age (X1)	edu (X2)	score (C)
Ann	**F**	[30,35)	BS	**low**

S binary membership in a protected group (gender)

X "regular" attribute-value pairs (age and edu)

C binary classification outcome (score)

support $(S\ X \rightarrow C)$ = % D that assigns the same attribute values for S, X and C

confidence $(S\ X \rightarrow C)$ = support $(S\ X \rightarrow C)$ / support $(S\ X)$

α-protection $(S\ X \rightarrow C)$ = confidence $(S\ X \rightarrow C)$ / confidence $(X \rightarrow C)$

Homework 6

- Read about the ProPublica COMPAS investigation here https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing and here https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm
- Download the post-processed subset of the ProPublica COMPAS dataset from the course website, and load it into your PostgreSQL database on tux
- Write a sequence of SQL queries that takes as input support, confidence and protection thresholds and outputs all PD-CARs that pass the specified thresholds
    - assume that *race* is the protected attribute and that *v-decile* (medium or high violent decile score) is the outcome
    - assume that "regular" attributes are *gender*, *marriage* and *age*, and that *age* is discretized appropriately (you don't need to do any data post-processing)
    - use SQL code provided on the course website as your starting point, do not change the format of the output

Data, responsibly

Because of its tremendous **power**, massive data analysis must be used **responsibly**

Fairness

Diversity

Transparency

Data analysis in context

- A touch of probability and statistics
  - Events and outcomes
  - Independent and non-independent events
  - Avoiding the pitfalls: the 4 Cs, Bonferroni principle and Oakham's razor

Probability and Statistics

- Experiments have outcomes
- An event is an occurrence of a set of outcomes
- Probability of an event measures how likely this event is to occur

Example: tossing a fair coin, on Earth, subject to the usual gravity laws

Experiment: toss the coin once

Outcomes: {H, T}

Events: {H}, {T}, {} - neither heads nor tails, {H,T} - either heads or tails

Probabilities of events: P({H}) = 0.5 P({T}) = 0.5, P{{}} = 0, P{{H,T}} = 1

Probabilities of disjoint events: P({H}) = 0.5 P({T}) = 0.5

What changes if the coin is biased?

What if our experiment is to toss the coin twice?

Independent Events

Events A and B are independent if the fact that A occurs does not affect the probability of B occurring.

Experiment: toss 2 fair coins, call them A and B.

We write A=1 if A lands heads, A=0 if A lands tails.

Outcomes: {A=0, B=0}, {A=0, B=1}, {A=1, B=0}, {A=1, B=1}

$P(A) = 1/2 \quad P(B) = 1/2$

$P(A, B) = 1/4$

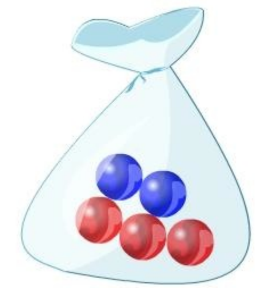
More than 2 independent events?

$P(A, B) = P(A)\,P(B)$

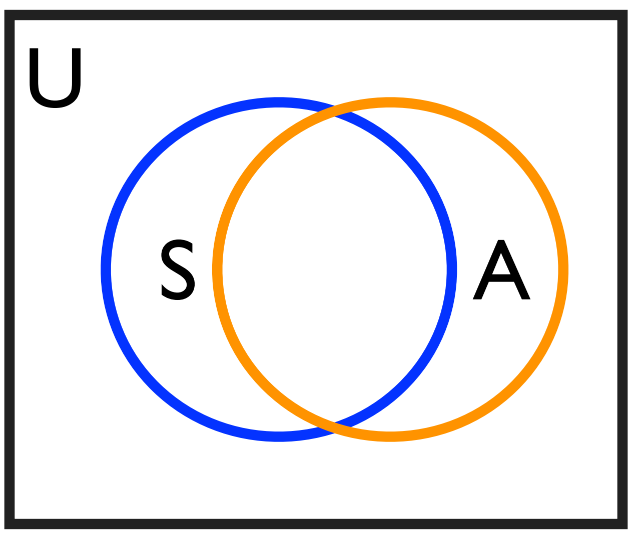
Conditional probability

Conditional probability measures the probability of an event given that another event has occurred.

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$



Example: S - probability that a person smokes; A - probability that there is an ashtray in the person's home



| U | = 100 people

| S | = 15 smokers

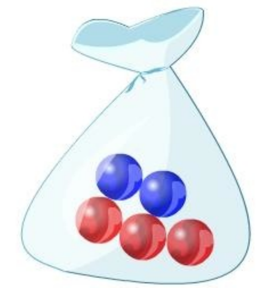
| A | = 20 have ashtrays in their home

| A and S | = 10

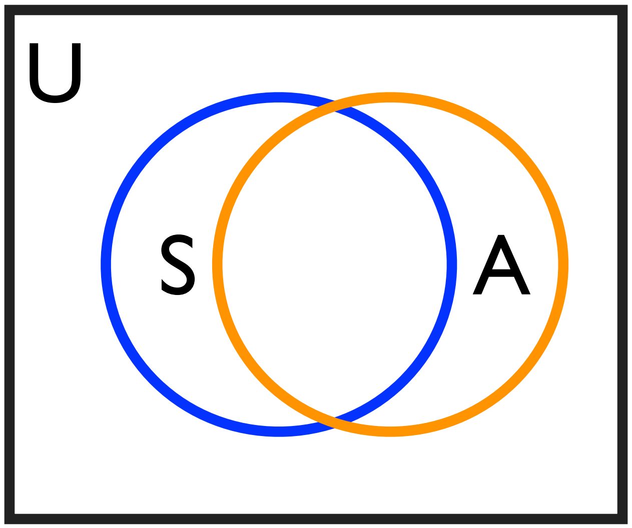
Conditional probability

Conditional probability measures the probability of an event given that another event has occurred.

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$



Example: S - probability that a person smokes; A - probability that there is an ashtray in the person's home



P(S) = 0.15   P(A) = 0.2

P(S and A) = 0.1

P(S | A) = 0.1 / 0.2 = 0.5

P(A | S) = 0.1 / 0.15 = 0.67

Baye's rule: $P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$

Co-occurrence

Co-occurence: events A and B *frequently* occur together

An event is an assignment of a value to a variable (age, edu, income, astrological sign)

An experiment is a particular tuple in the database

Outcomes:
$$dom(A) \times dom(E) \times dom(I) \times dom(S)$$

age	edu	inc	sign
20	BS	27K	Aries
35	MS	102K	Taurus
62	PhD	200K	Virgo
70	BS	80K	Leo
…	…	…	…

P(sign = Taurus AND edu = BS) = 5%

is 5% frequent?

P(sign = Taurus AND edu = HS) = 2%

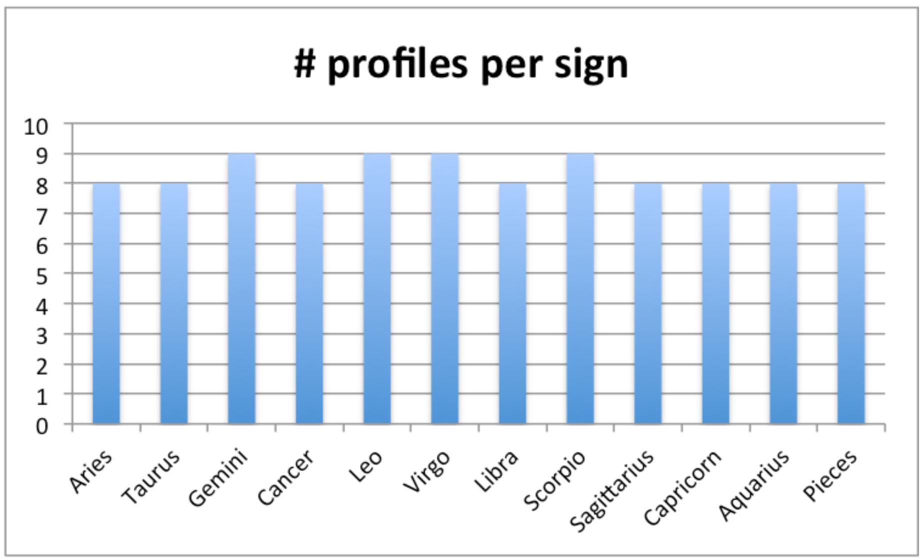
is 2% frequent?

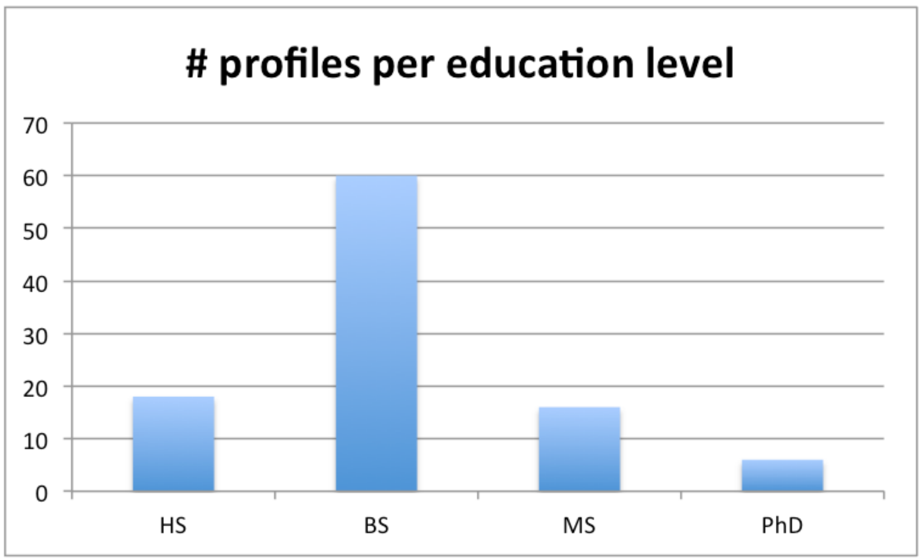
Co-occurrence

Co-occurence: events A and B *frequently* occur together

P(sign = Taurus AND edu = BS) = 5% *is 5% frequent?*

P(sign = Taurus AND edu = HS) = 2% *is 2% frequent?*



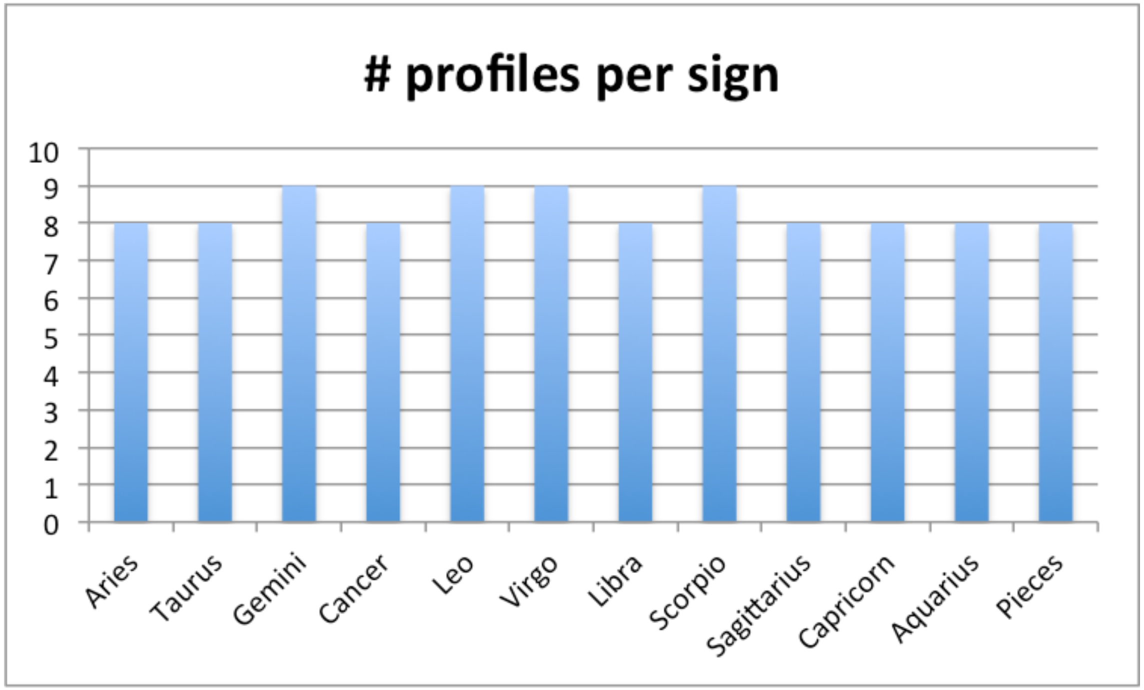


Co-occurrence vs. correlation

Co-occurence: events A and B *frequently* occur together

Correlation: events A and B occur together *more frequently than by random chance*

P(sign = Taurus) = 8%



Considerations of model and sample size!

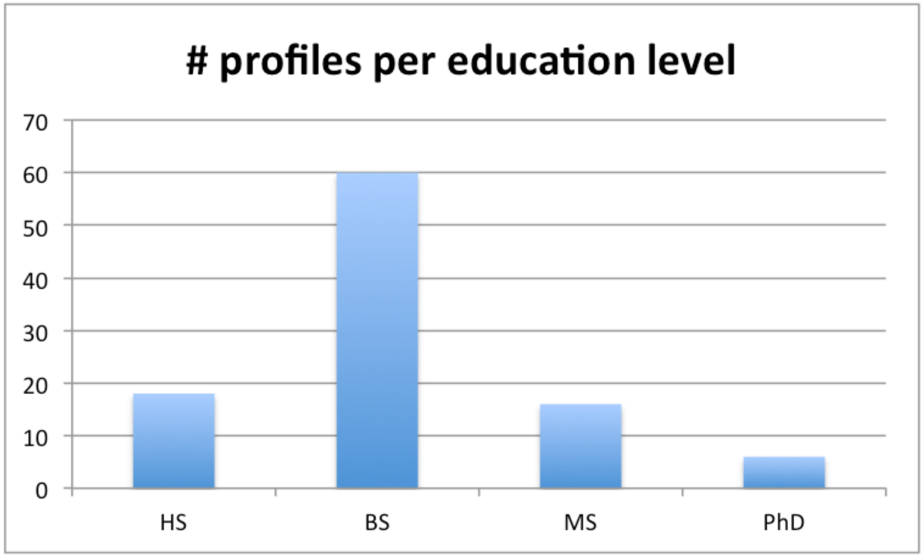
Co-occurrence vs. correlation

Co-occurence: events A and B *frequently* occur together

Correlation: events A and B occur together *more frequently than by random chance*

P(sign = Cancer AND edu = PhD) = 2%

P(sign = Libra AND edu = PhD) = 0



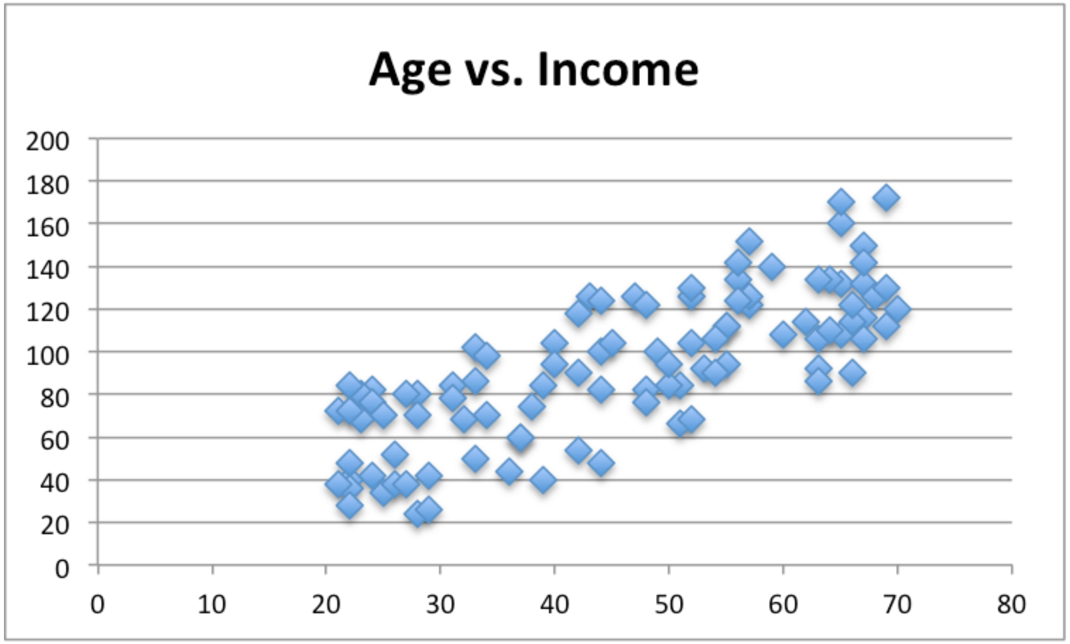
Considerations of model and sample size!

Correlation vs. causation

Co-occurence: events A and B *frequently* occur together

Correlation: events A and B occur together *more frequently than by random chance*

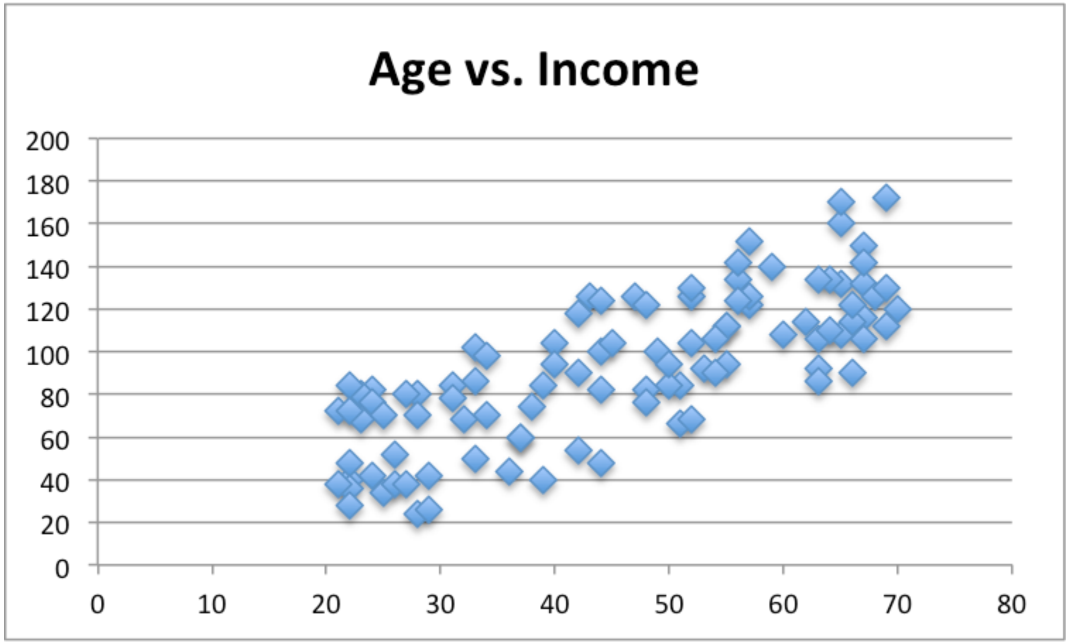
Causation: if event A occurs, then event B is *more likely to occur*



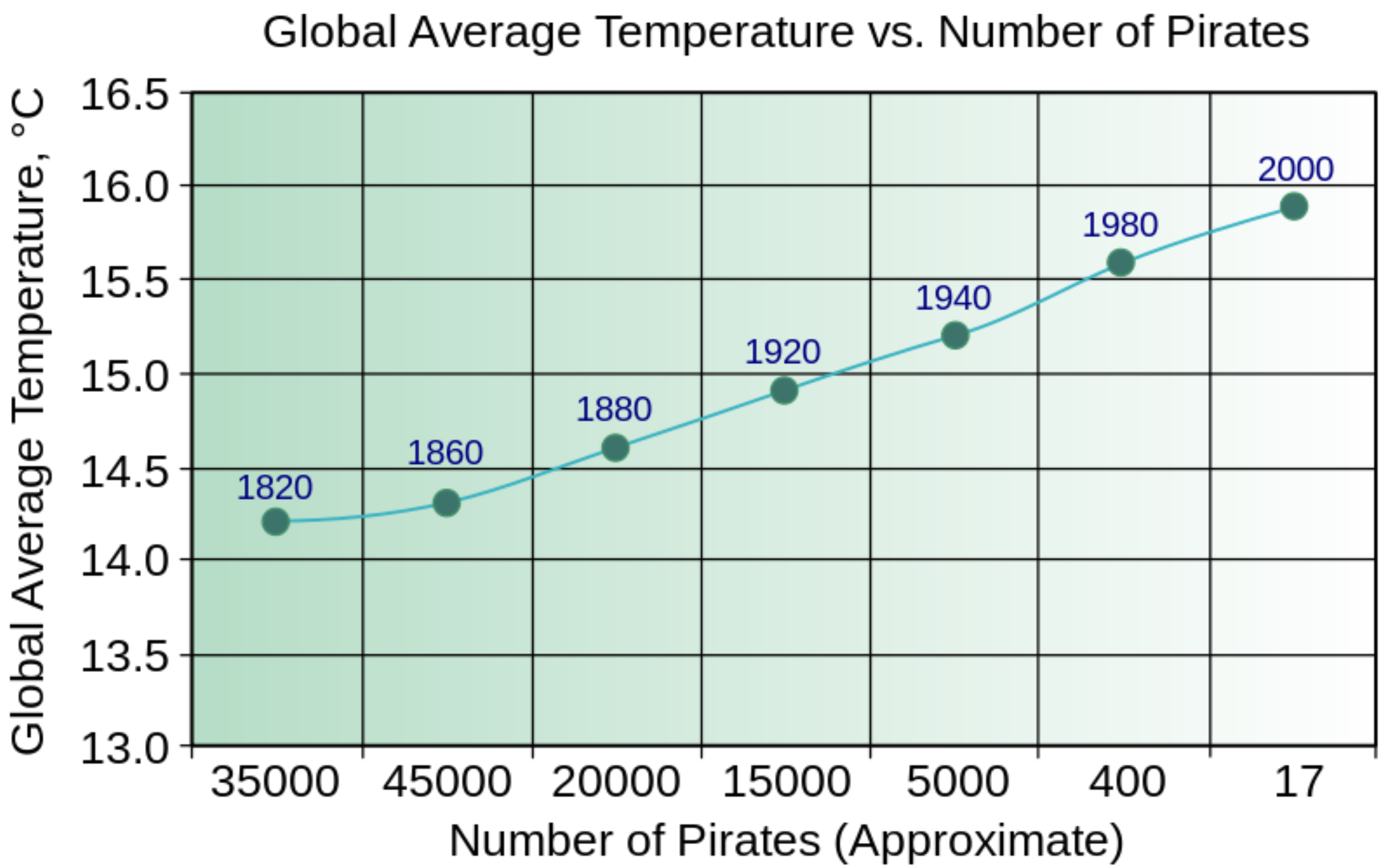
Correlation vs. causation

For correlated events A and B, the following relationships are possible

1. A causes B (direct causation)
2. B causes A (reverse causation)
3. A and B are consequences of a common cause
4. A causes B and B causes A (cyclic causation)



Causation vs. coincidence



see also http://www.tylervigen.com/spurious-correlations
for lots of fun examples

Bonferroni principle

In massive datasets, unusual events may appear, by coincidence, to be more frequent than expected. If the expected frequency of an event is lower than what is expected by random chance - the event cannot be reliably detected!

Example (sec 1.2.3 of LRU): “evil-doers” gather at hotels to plan evil deeds. A pair of people who is at the same hotel on 2 or more occasions are potentially evil-doers.

There are 10^{12} people (1 billion). Everyone decides with probability 0.01 to stay at a random hotel on any given day. There are 10^5 hotels. We examine 1000 days worth of hotel records.

What is the probability that 2 people will visit the same hotel on 2 separate occasions?

$$5 \times 10^{17} \times 5 \times 10^{5} \times 10^{-18} = 250{,}000$$

Oakham’s razor

Lex parsimoniae

If multiple hypotheses explain an observation, the simplest one should be preferred.

Used as a heuristic to help identify a promising hypothesis to test

Ockham’s motivation: can one prove the existence of God?

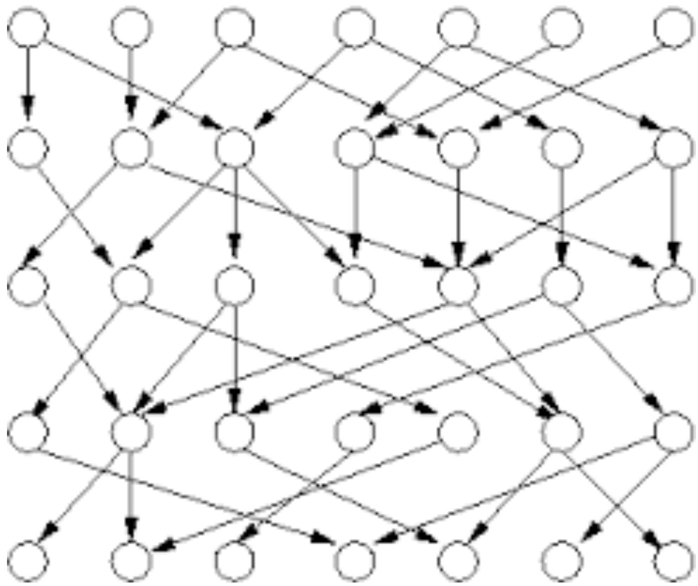
Many applications today: biology, probability theory, ethics

William of Ockham (1285-1347)

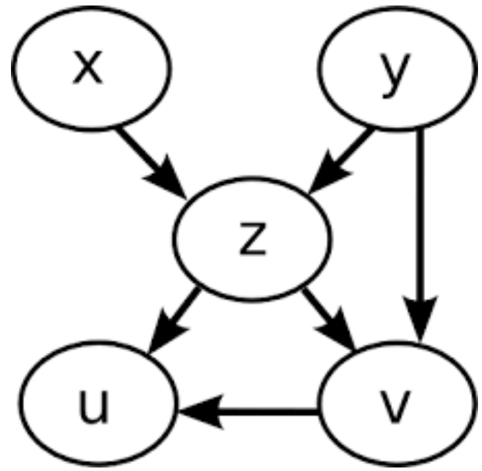
An example: the art of medical diagnosis

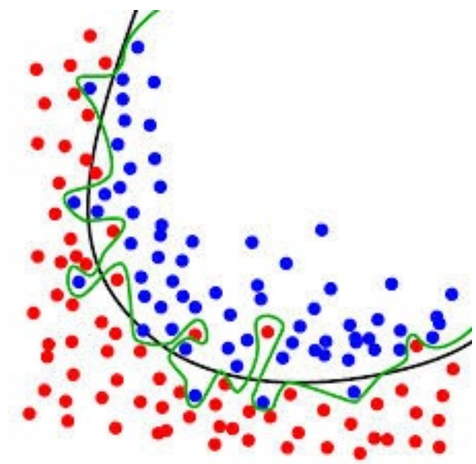
"It's your ribs. I'm afraid they're delicious." - New Yorker Cartoon
By: Paul Noth Item #: 10684223

An example: overfitting

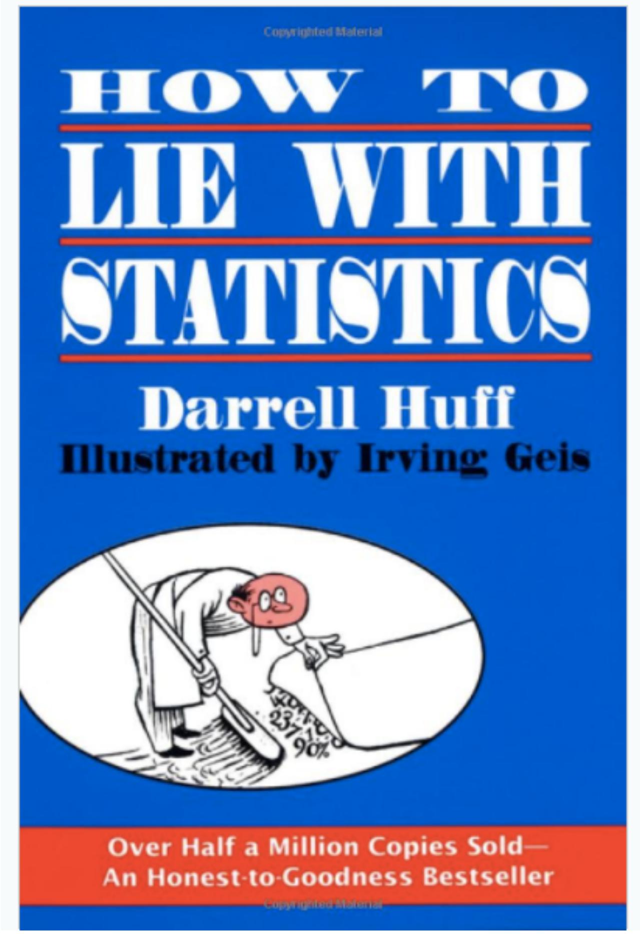


vs.





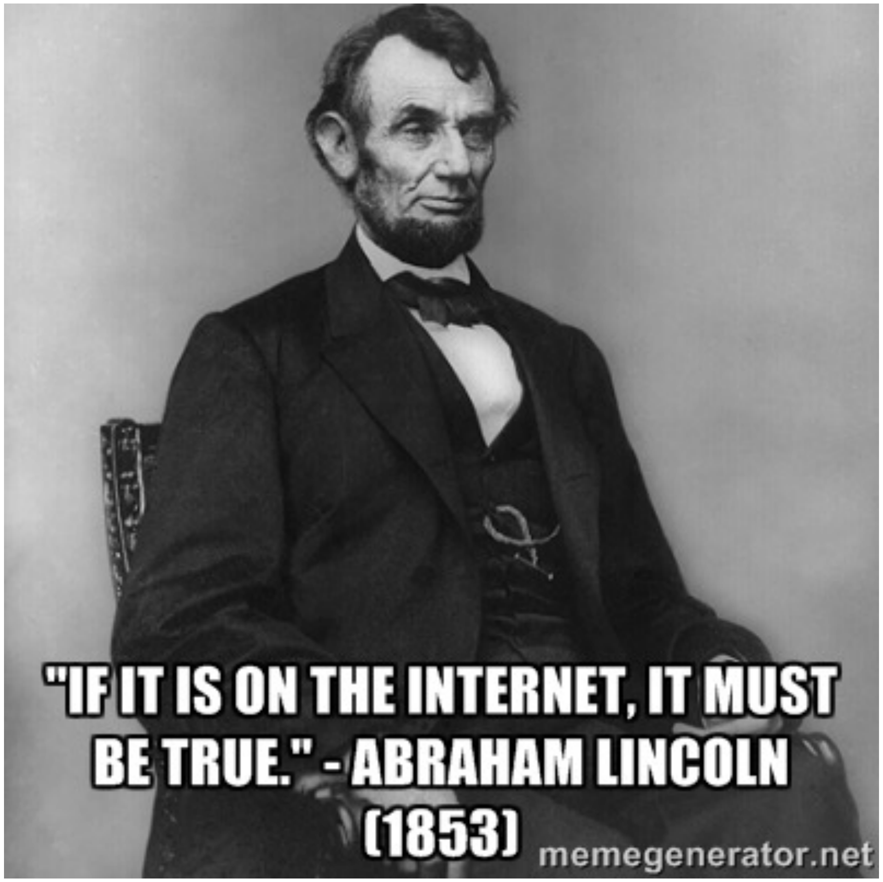
How to lie with Big Data



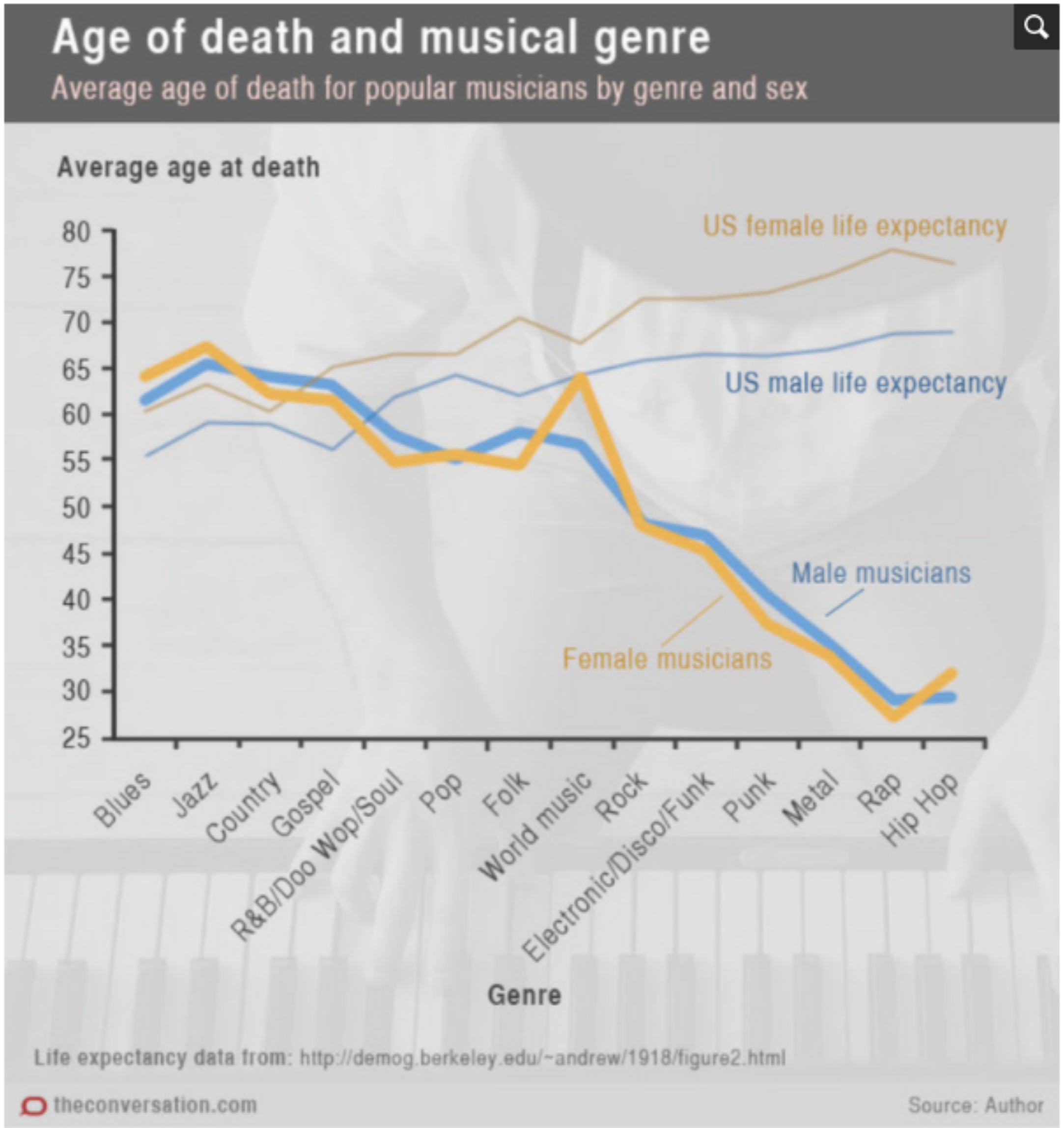
Statistics scares people, big data REALLY scares people!

Learn to question!

- Concepts
  - **understanding** data acquisition methods and data analysis processes
  - **verifying** the data and the process: provenance, credit attribution, trust
  - **interpreting** results



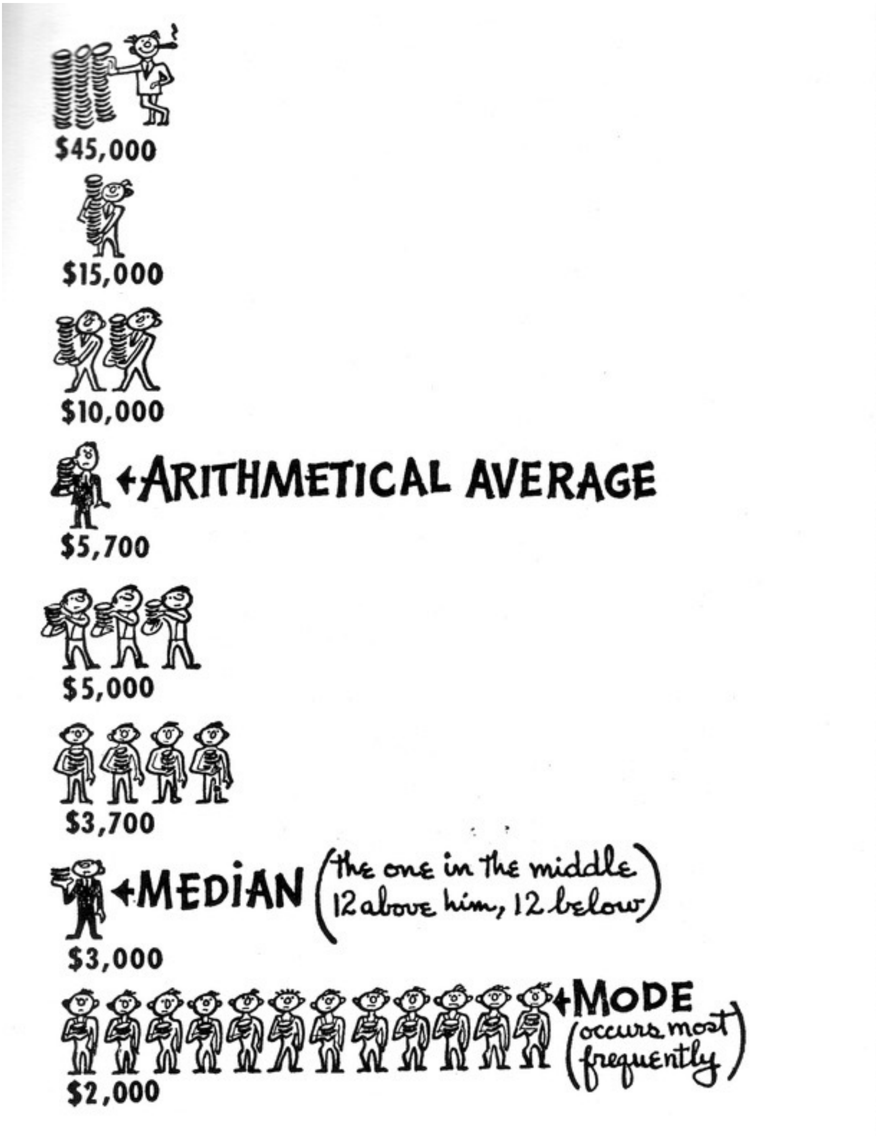
Case study: musicians and mortality

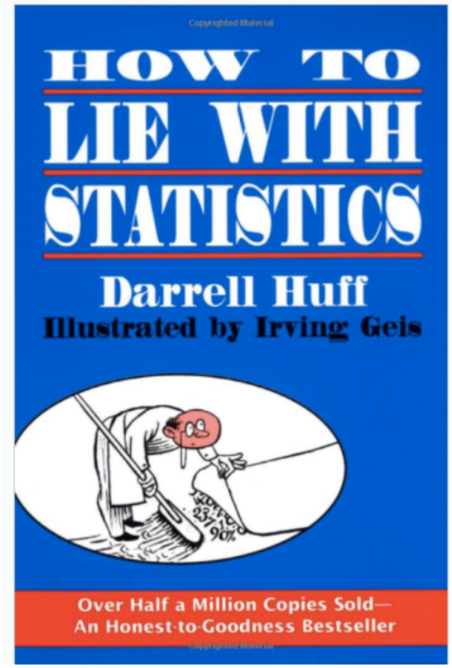


https://theconversation.com/music-to-die-for-how-genre-affects-popular-musicians-life-expectancy-36660

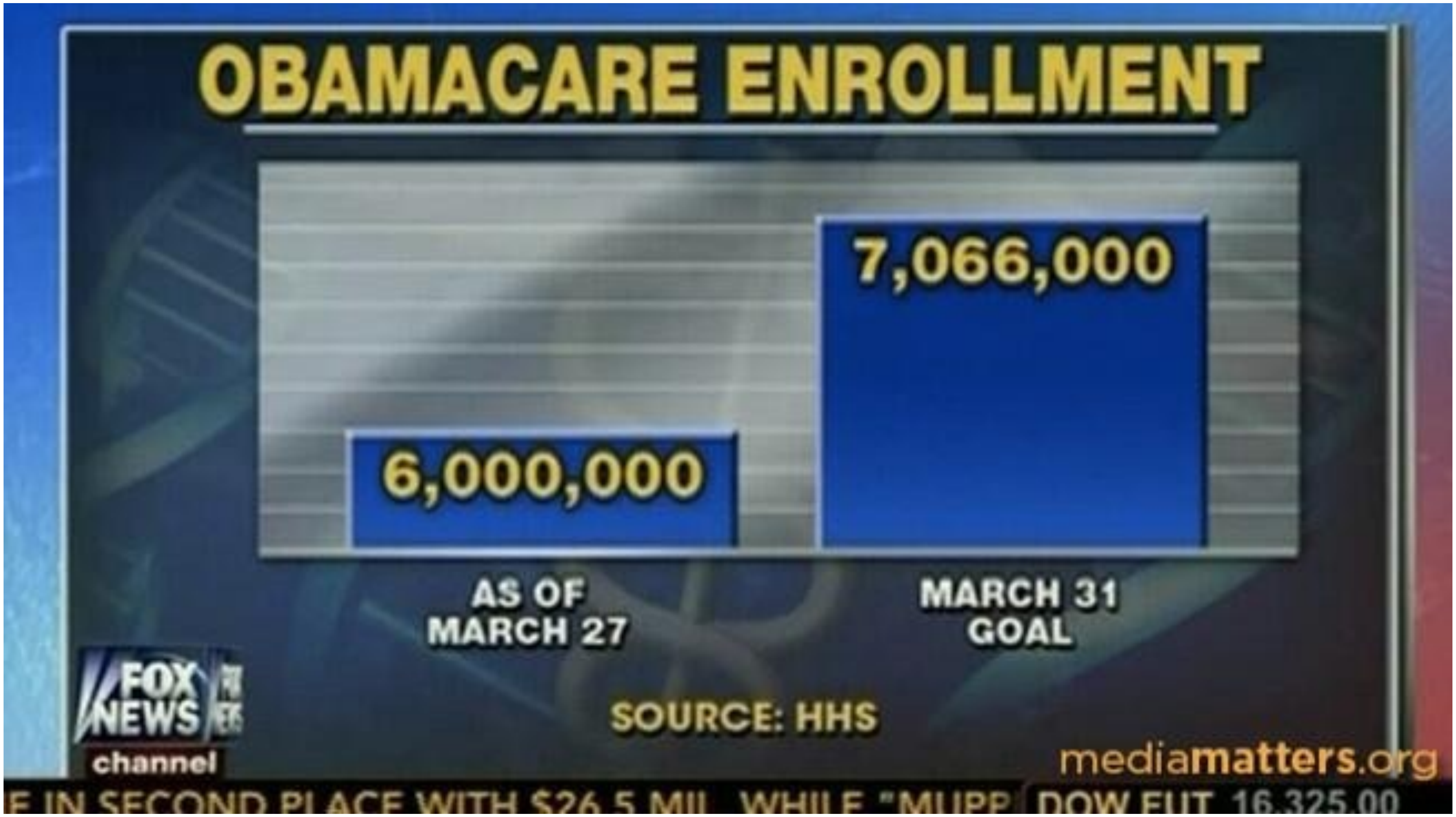
callingbullshit.org/case_studies

A well-chosen average



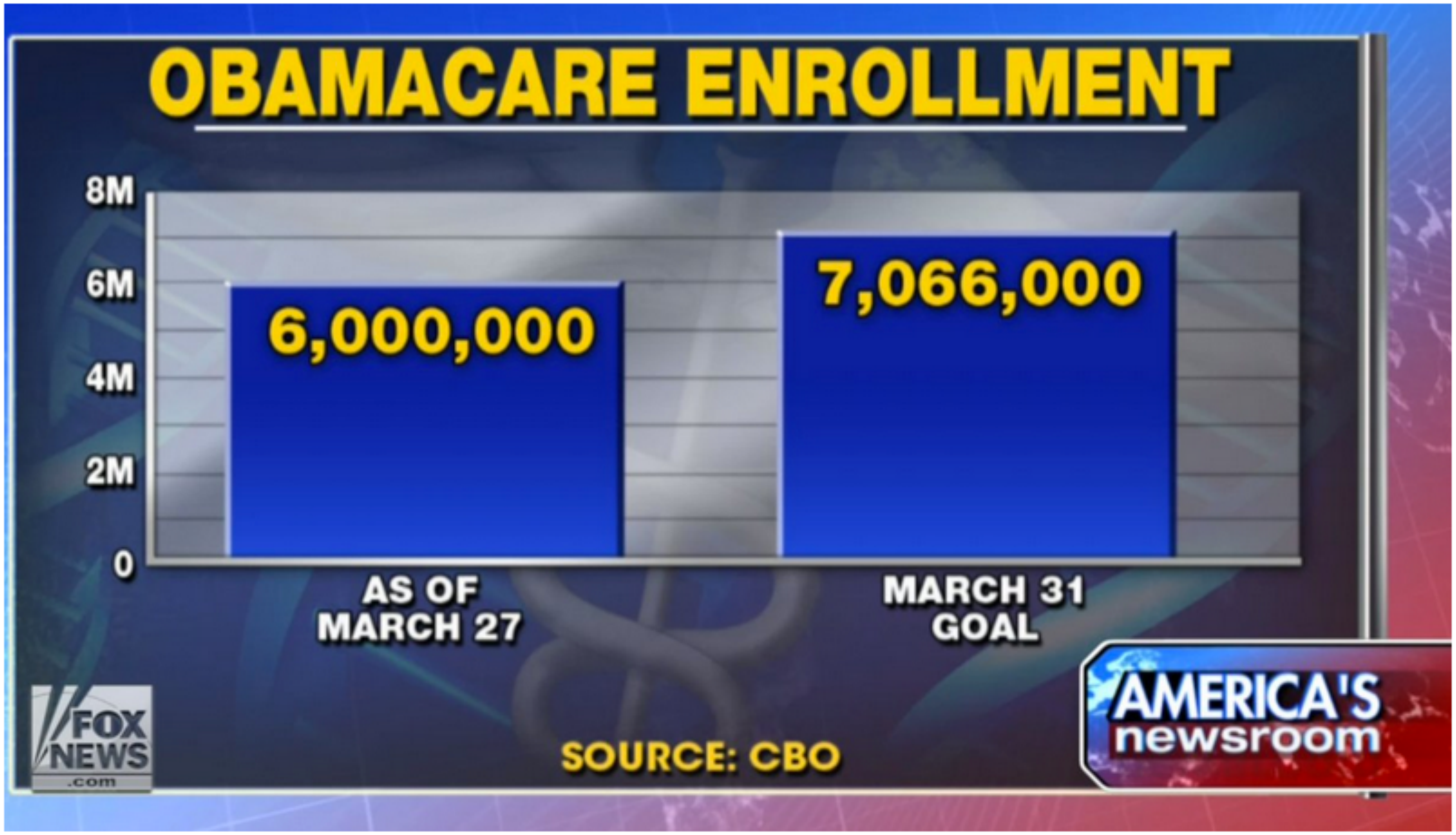


Case study: Fox news on Obamacare



http://www.businessinsider.com/fox-news-obamacare-chart-2014-3

Case study: Fox news on Obamacare



Fox News

http://www.businessinsider.com/fox-news-obamacare-chart-2014-3

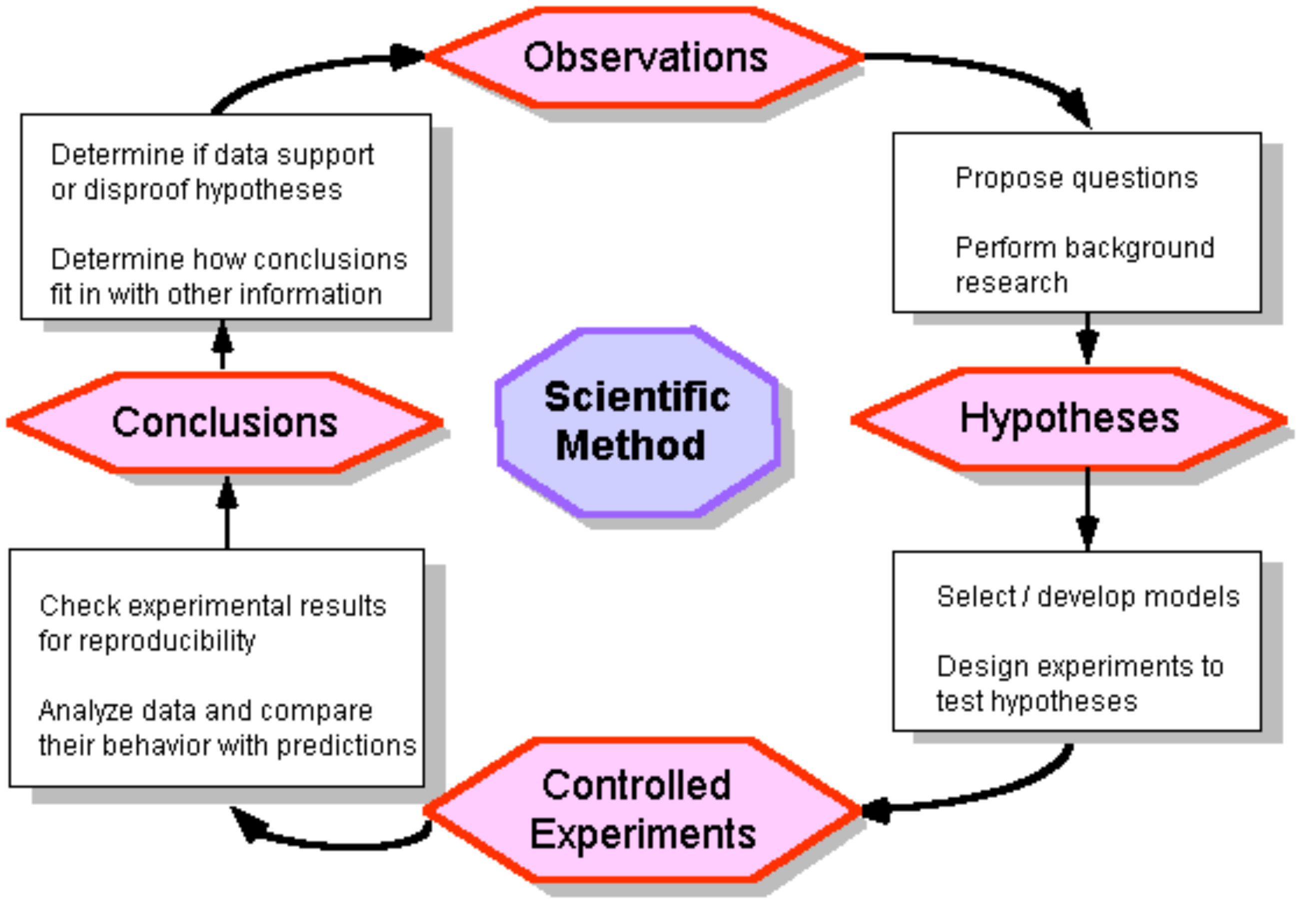
The New York Times

The truth is more important now than ever.

Watch the video

https://www.nytimes.com/subscriptions/Multiproduct/lp3L3QR.html?campaignId=6L9HJ

Is data science a science?



http://bioserv.fiu.edu/~walterm/human_online/labs/scientific_meth/sci_meth1/scientific_method_files/image001.gif

The scientific method

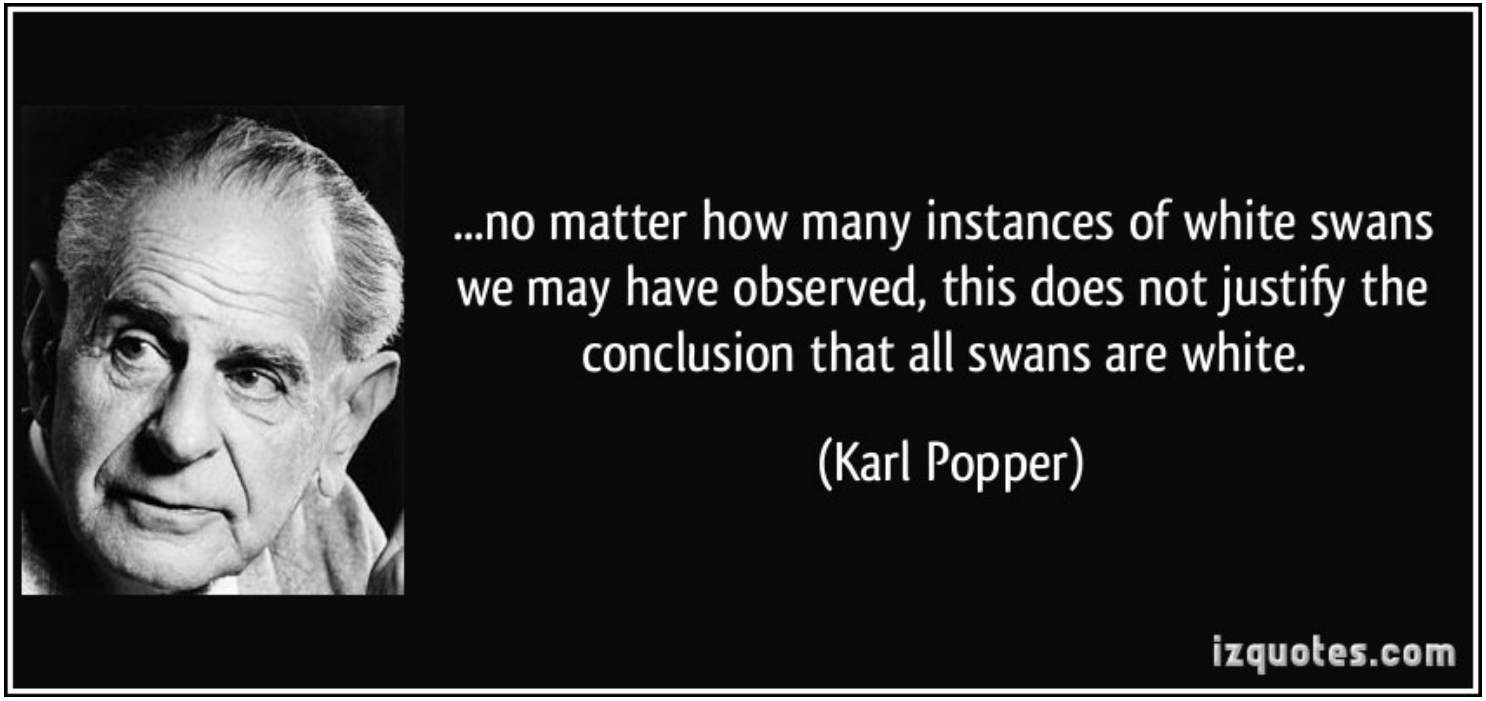
A method or procedure that has characterized natural science since the 17th century, consisting in systematic observation, measurement, and experiment, and the formulation, testing, and modification of hypotheses.

The Oxford English Dictionary

- hypothesis is formulated based on observation
- hypothesis is tested under controlled conditions using sound reasoning
- avoids confirmation bias - people tend to observe what they expect to observe
- the process is reproducible

Falsifiability

A crucial component of the scientific method - every hypothesis must be falsifiable (refutable, testable), i.e., it must be possible to prove that the statement in question is false



Is astrology a science?

http://www.bodymemory.com/uploads/2/3/2/8/23288628/
s266135201567006809_p25_i2_w1000.jpeg

Data, responsibly

Because of its tremendous **power**, massive data analysis must be used **responsibly**

Fairness Diversity Transparency

let's look at diversity next

Illustration: online dating

Dating query: female, 40 or younger, at least some college, in order of decreasing income

Results are homogeneous at top ranks

Both the seeker (asking the query) and the matches (results) are dissatisfied

Crowdsourcing, crowdfunding, ranking of Web search results, … - all subject to this problem

the rich get richer, the poor get poorer


MBA, 40 years old
makes $150K


MBA, 40 years old
makes $150K


MBA, 40 years old
makes $150K


MBA, 40 years old
makes $150K

… 999 matches


PhD, 36 years old
makes $100K

… 9999 matches


BS, 27 years old
makes $80K

Rank-aware clustering


MBA, 40 years old
makes $150K


MBA, 40 years old
makes $150K


MBA, 40 years old
makes $150K


MBA, 40 years old
makes $150K

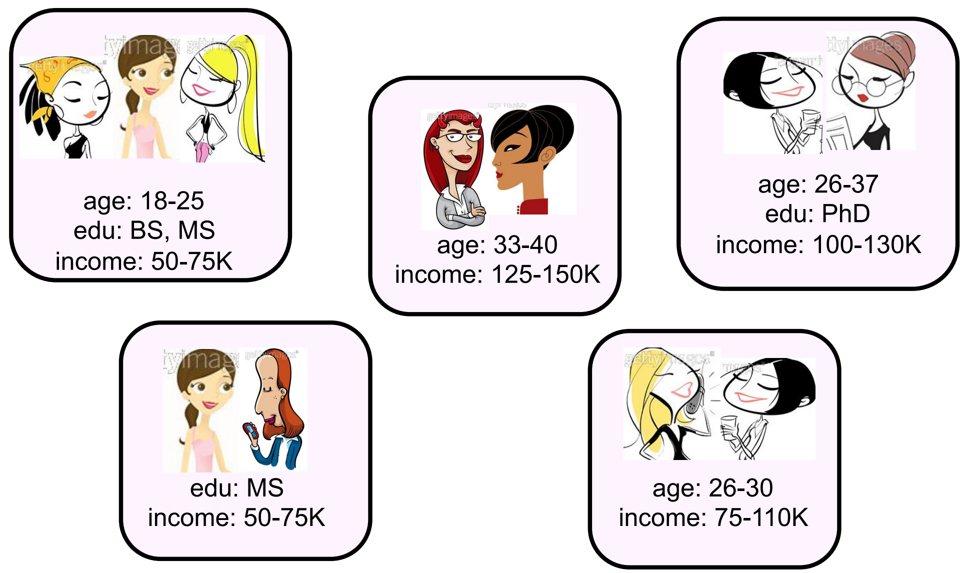
… 999 matches


PhD, 36 years old
makes $100K

… 9999 matches


BS, 27 years old
makes $80K

Return clusters that expose **best from among comparable** items (profiles) w.r.t. user preferences



More diverse items seen, and liked, by users

Users are more engaged with the system

Diversity & friends

- For a given user consuming information in search and recommendation, relevance is important, but so are:
  - **diversity** - avoid returning similar items
  - **novelty** - avoid returning known items
  - **serendipity** - surprise the user with unexpected items
- For a set of users
  - uncommon information needs must be met: **less popular** "in the tail" queries constitute the overwhelming majority
  - lack of diversity can lead to **exclusion**

> **Jonas Lerman:** "… the nonrandom, systematic omission of people who live on big data's margins, whether due to poverty, geography, or lifestyle…"

Diversity when data is about people

- Data must be **representative** - bias in data collection may be amplified in data analysis, perpetuating the original bias

- In this sense diversity is related to **coverage**

Data, responsibly

Because of its tremendous **power**, massive data analysis must be used **responsibly**

Fairness

Diversity

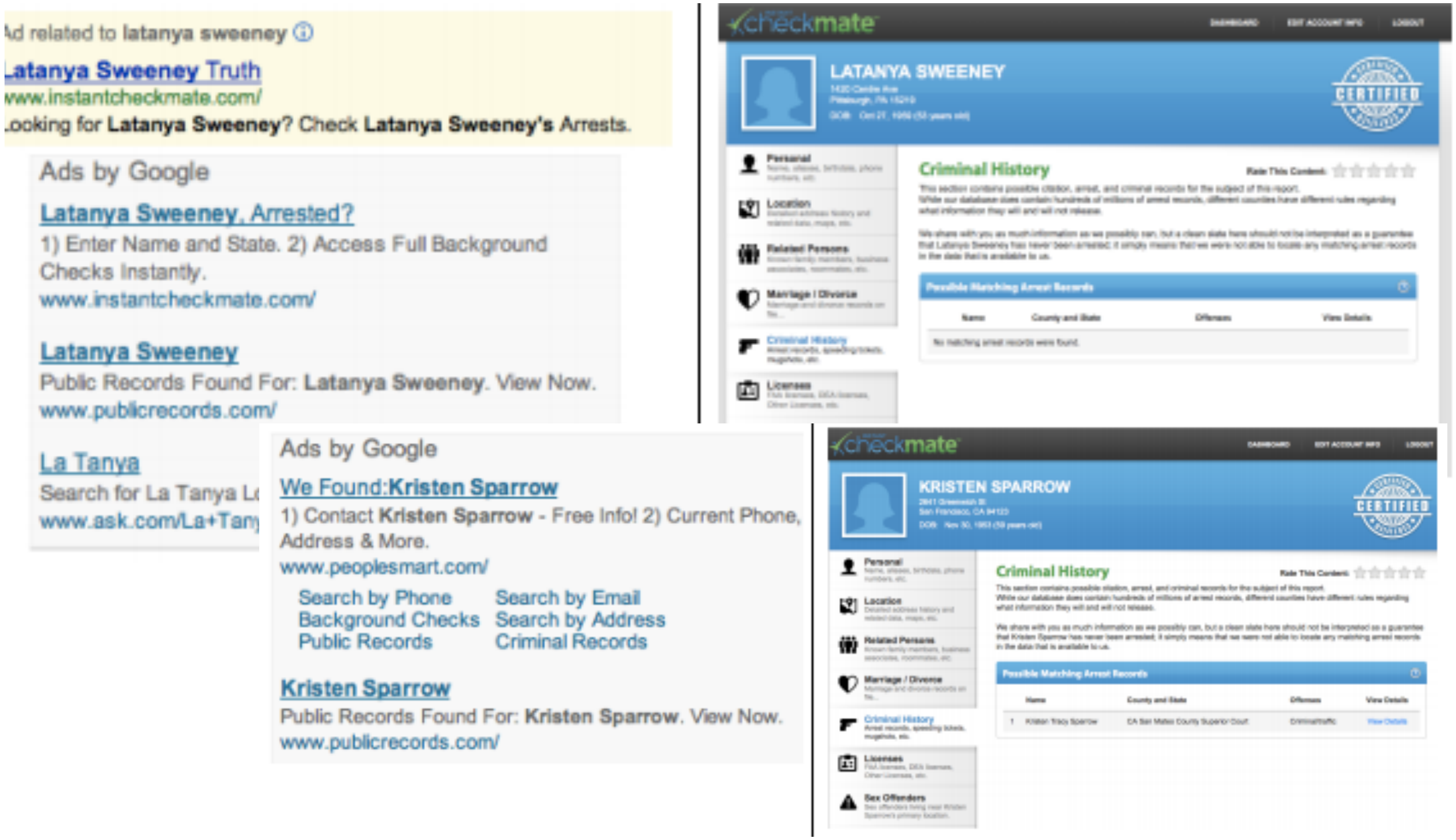
Transparency

and now transparency

Racially identifying names

[Latanya Sweeney; *CACM 2013*]

Google AdSense



racially identifying names trigger ads suggestive of an arrest record

Transparency and accountability

- Users and regulators must be able to **understand** how raw data was selected, and what operations were performed during analysis
- Users want to **control** what is recorded about them and how that information is used
- Users must be able to **access** their own information and correct any errors (US Fair Credit Reporting Act)
- **Transparency** facilitates **accountability** - verifying that a service performs as it should, and that data is used according to contract
- Related to **neutrality**, more on this later

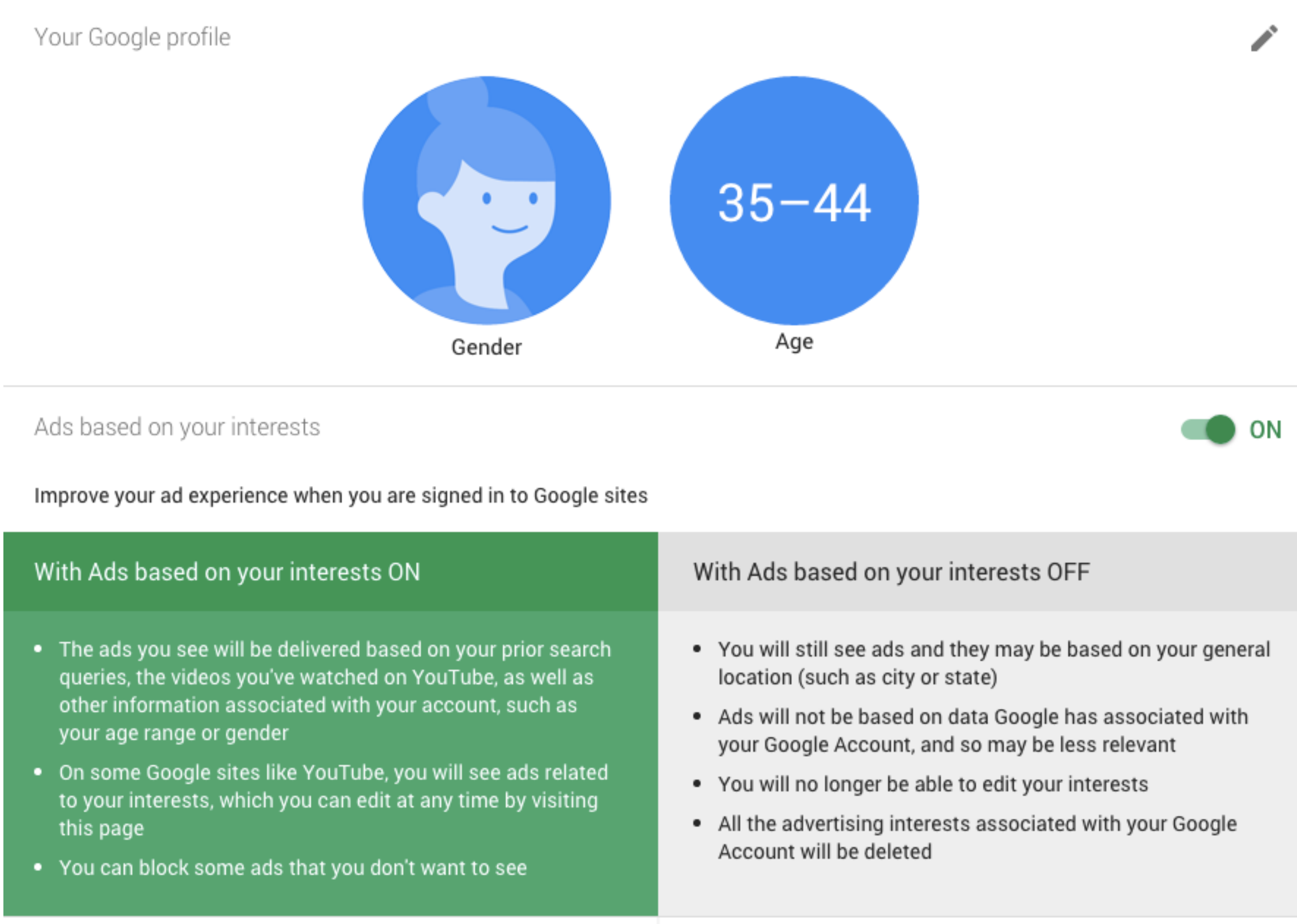
the problem is broad, we focus on a specific case

Example: Ad targeting online

- **Users** browse the Web, consume content, consume ads (see / click / purchase)
- **Content providers** outsource advertising to third-party ad networks, e.g., Google's DoubleClick
- **Ad networks** track users across sites, to get a global view of users' behaviors
- **Google Ad Settings** aims to provide **transparency** / give **control to users** over the ads that they see

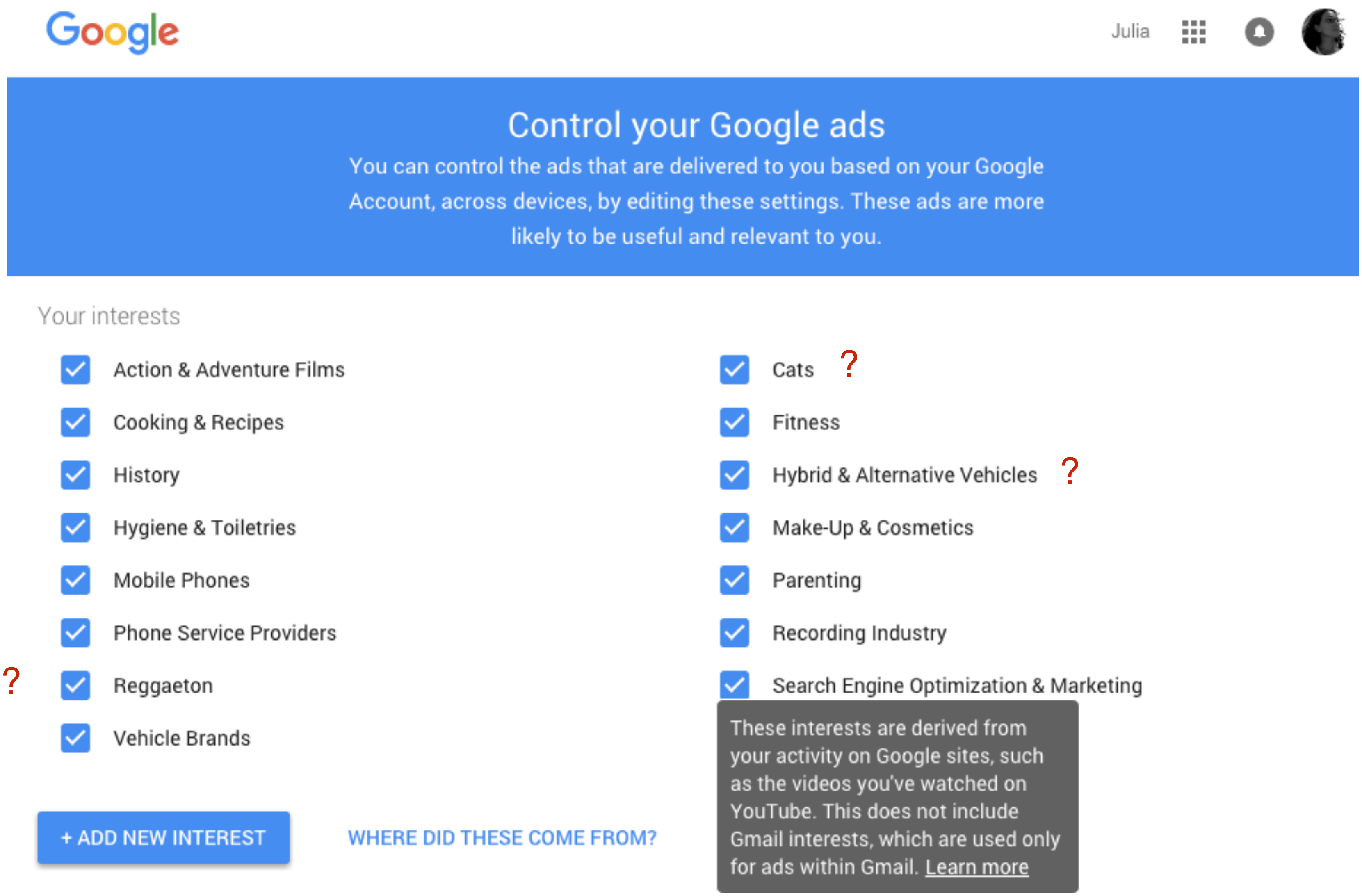
do users truly have transparency / choice or is this a placebo button?

Google Ads Settings



http://www.google.com/settings/ads

Google Ads Settings

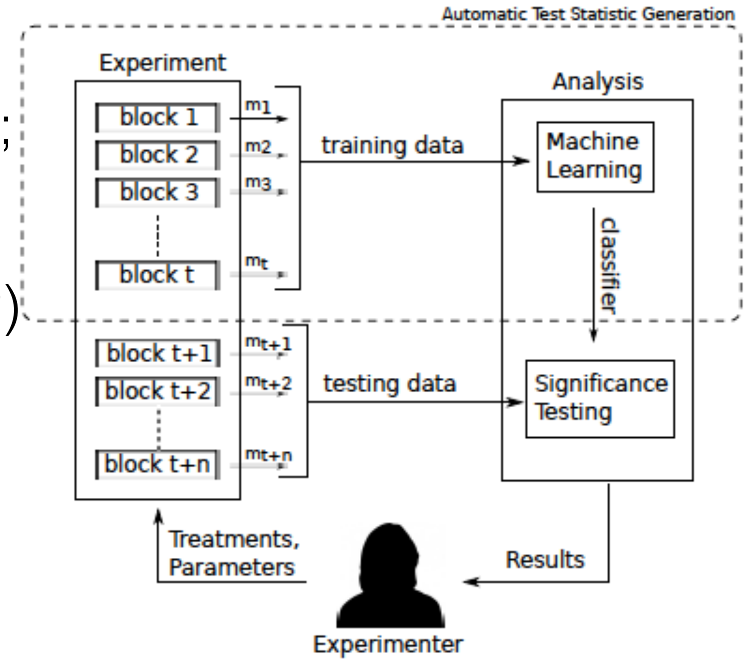


http://www.google.com/settings/ads

AdFisher: methodology

[Amit Datta, Michael C. Tschantz, Anupam Datta; *PETS 2015*]

- Browser-based experiments, simulated users
  - **input**: (1) visits to content providing websites; (2) interactions with Google Ad Settings
  - **output**: (1) ads shown to users by Google; (2) change in Google Ad Settings



- Fisher randomized hypothesis testing
  - **null hypothesis** inputs do not affect outputs
  - control and experimental treatments
  - AdFisher can help select a test statistic

AdFisher: discrimination

[Amit Datta, Michael C. Tschantz, Anupam Datta; *PETS 2015*]

Non-discrimination: Users differing only in protected attributes are treated similarly

Causal test: Find that a protected attribute changes ads

Experiment 1: **gender and jobs**

Specify gender (male/female) in Ad Settings, simulate interest in jobs by visiting employment sites, collect ads from Times of India or the Guardian

Result: males were shown ads for higher-paying jobs significantly more often than females (1852 vs. 318)

violation

AdFisher: transparency

[Amit Datta, Michael C. Tschantz, Anupam Datta; *PETS 2015*]

Transparency: User can view data about him used for ad selection

Causal test: Find attribute that changes ads but not settings

Experiment 2: **substance abuse**

Simulate interest in substance abuse in the experimental group but not in the control group, check for differences in Ad Settings, collect ads from Times of India

Result: no difference in Ad Settings between the groups, yet significant differences in what ads are served: rehab vs. stocks + driving jobs

violation

AdFisher: accountability

[Amit Datta, Michael C. Tschantz, Anupam Datta; *PETS 2015*]

Ad choice: Removing an interest decreases the number of ads related to that interest.

Causal test: Find that removing an interest causes a decrease in related ads

Experiment 3: **online dating**

Simulate interest in online dating in both groups, remove "Dating & Personals" from the interests on Ad Settings for experimental group, collect ads

Result: members of experimental group do not get ads related to dating, while members of the control group do

compliance

Power comes with responsibility

power

A handful of big players command most of the world's computational resources and most of the data, including all of your personal data - an **oligopoly** (def: a state of limited competition, in which a market is shared by a small number of producers or sellers)

danger

can destroy business competition

control what information you receive

can guide your decisions

can infringe on your privacy and freedom

Additional information and resources

- "How to lie with statistics", Darrell Huff, 1954
- "Weapons of math destruction: How Big Data increases inequality and threatens democracy", Cathy O'Neil, 2016
- Calling bullshit in the age of Big Data callingbullshit.org
- The Data, Responsibly manifesto: http://wp.sigmod.org/?p=1900
- EDBT 2016 tutorial: https://www.cs.drexel.edu/~julia/documents/DataResponsibly.pdf
- Transparency in ranking: https://freedom-to-tinker.com/?p=12189&preview=true
- Fairness in ranking: https://arxiv.org/abs/1610.08559
- Diversity in ranking: https://www.cs.drexel.edu/~julia/documents/barac.pdf