CS102: Big Data

Tools and Techniques, Discoveries and Pitfalls

Spring 2017

Ethan Chan, Lisa Wang

Python and Pandas;

Correlation and Causation; Privacy

Announcements

- HW2 SQL due this Sunday
  - Use single quotes ‘’ for SQL assignment, and not double quotes “”
  - **1.0** * NUMSUCCESS / NUMTOTAL
- If still on the waitlist, speak to me after class

Now that you've learnt Python…

Welcome to Slithereen House

```
IMPORT OS
CUR_PATH = OS.GETCWD()
IGNORE_SET = SET(["__INIT__.PY",
"COUNT_SOURCELINES.PY"])
```

YEAH, HE'S A PARSER-TONGUE.
I DIDN'T KNOW HARRY SPOKE PYTHON.

Learning Goals

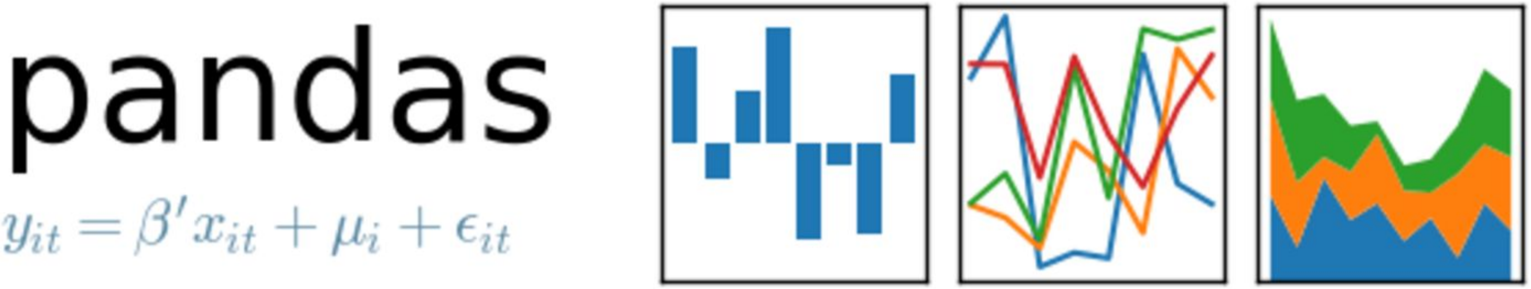
- Data Operations / Plotting in Python
- Data Operations in Pandas package
- Determine Correlation
- Determine if Causation
- Privacy

SQL vs Python

Python offers more complex functionality

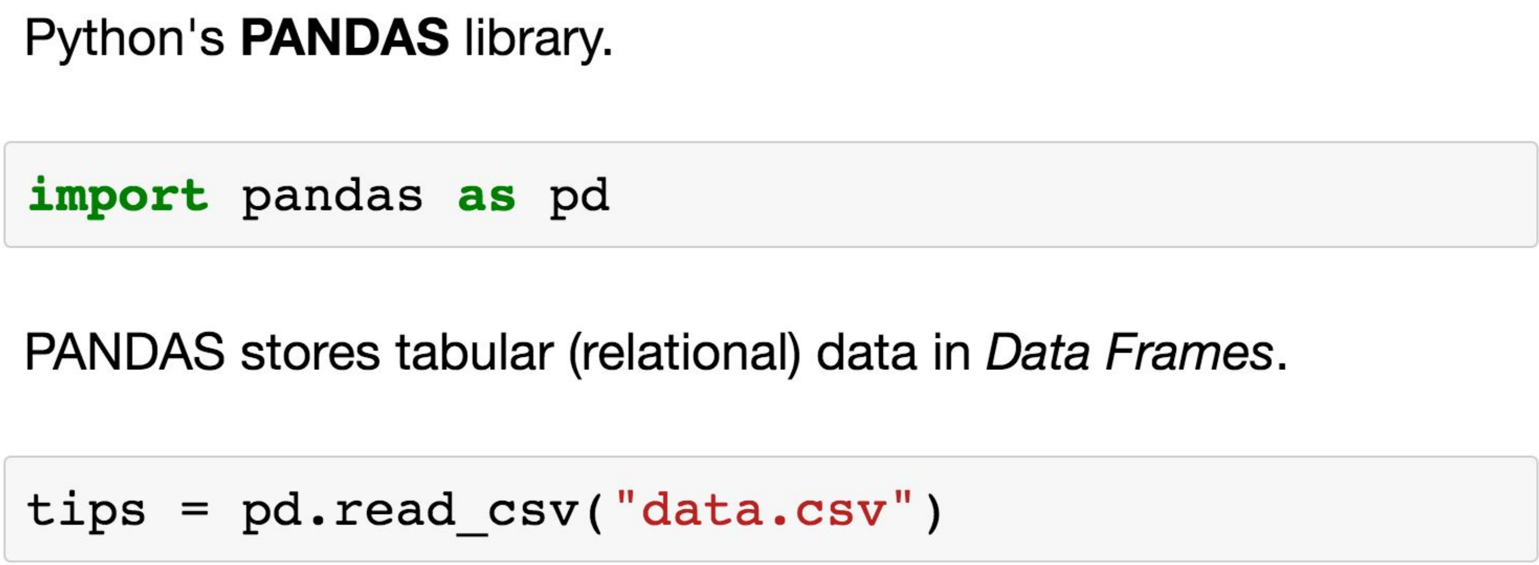
- Data exploration and maniplation a lot easier
- Packages such as pandas and scipy for data analysis
- Plotting

Use SQL for simple operations on large databases.
Use Python for more complex analyses.



- Open source library for data analysis written in Python language
- Stores tabular (relational) data in **Data Frames**

INITIALIZING A PANDAS DATAFRAME



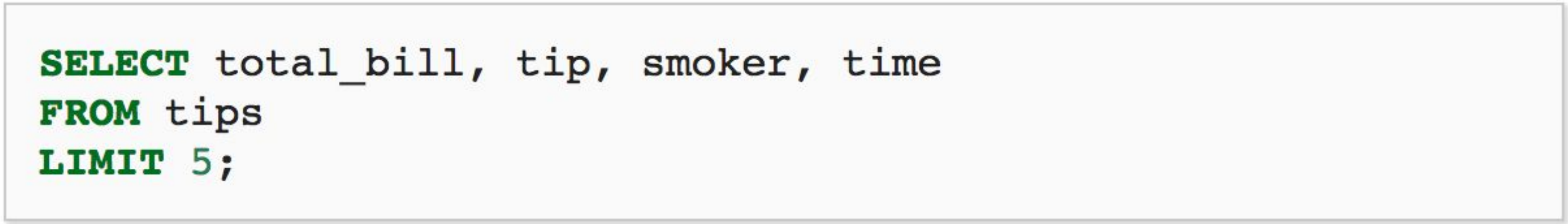
Python's **PANDAS** library.

```
import pandas as pd
```

PANDAS stores tabular (relational) data in *Data Frames*.

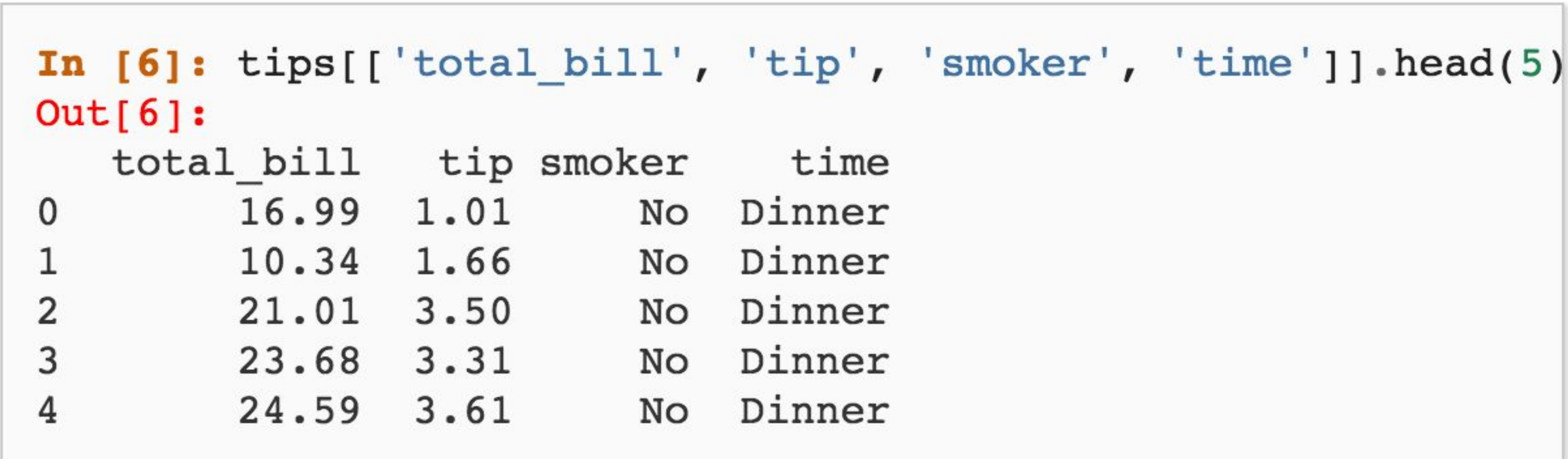
```
tips = pd.read_csv("data.csv")
```

SELECT - SQL vs Pandas comparison



```sql
SELECT total_bill, tip, smoker, time
FROM tips
LIMIT 5;
```

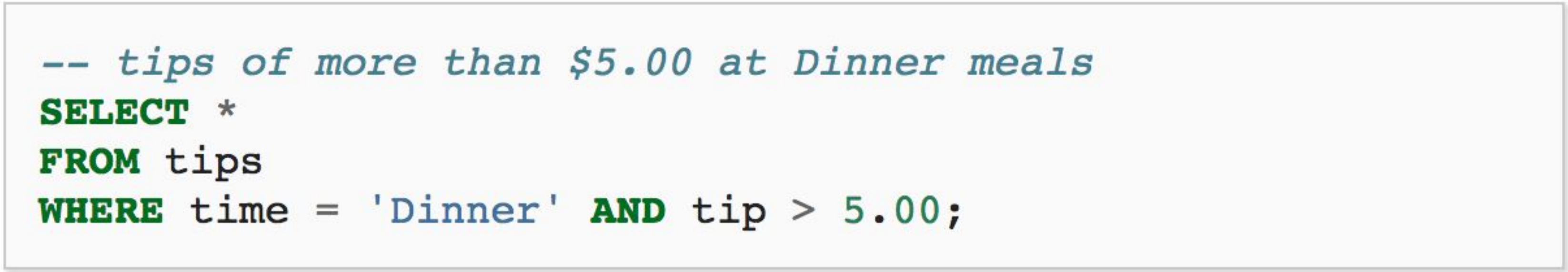
With pandas, column selection is done by passing a list of column names to your DataFrame:



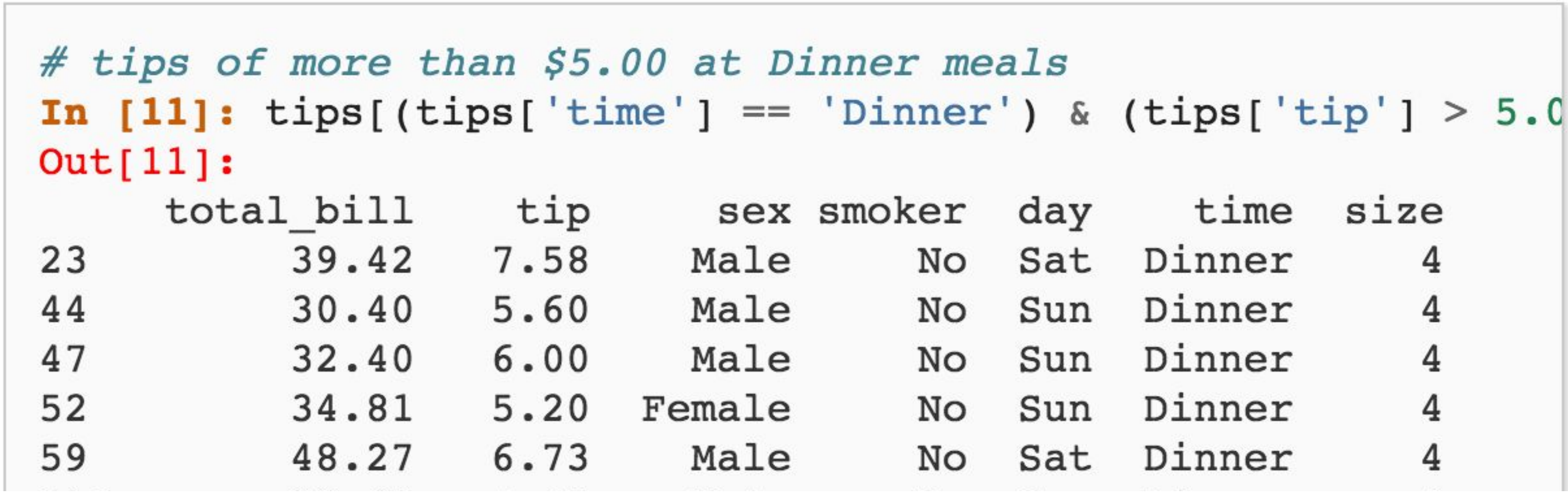
```
In [6]: tips[['total_bill', 'tip', 'smoker', 'time']].head(5)
Out[6]:
   total_bill   tip smoker    time
0       16.99  1.01     No  Dinner
1       10.34  1.66     No  Dinner
2       21.01  3.50     No  Dinner
3       23.68  3.31     No  Dinner
4       24.59  3.61     No  Dinner
```

source: pandas website

WHERE - SQL vs Python Comparison



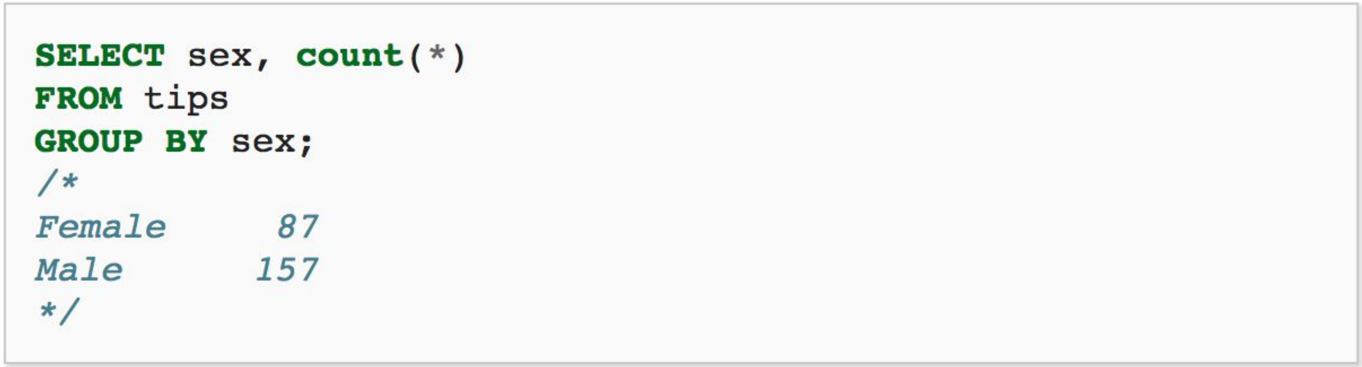
```
-- tips of more than $5.00 at Dinner meals
SELECT *
FROM tips
WHERE time = 'Dinner' AND tip > 5.00;
```



```
# tips of more than $5.00 at Dinner meals
In [11]: tips[(tips['time'] == 'Dinner') & (tips['tip'] > 5.0
Out[11]:
      total_bill   tip     sex smoker  day    time  size
23         39.42  7.58    Male     No  Sat  Dinner     4
44         30.40  5.60    Male     No  Sun  Dinner     4
47         32.40  6.00    Male     No  Sun  Dinner     4
52         34.81  5.20  Female     No  Sun  Dinner     4
59         48.27  6.73    Male     No  Sat  Dinner     4
```

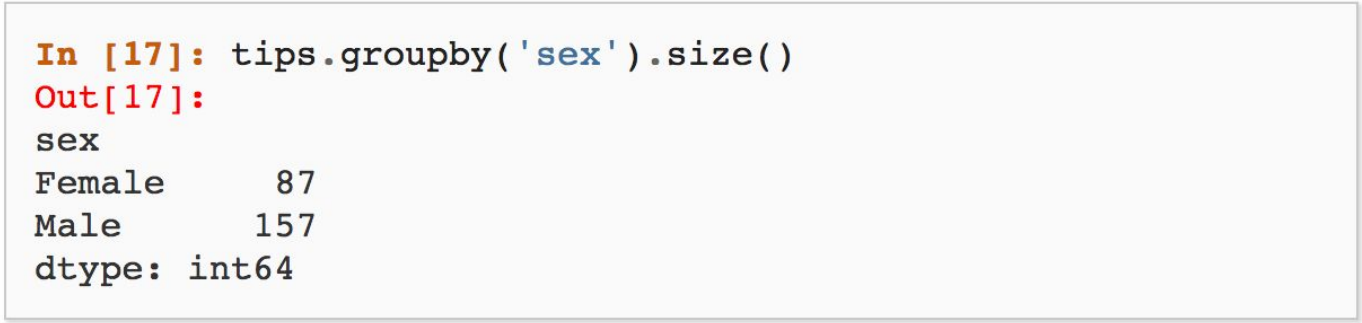
source: pandas website

GROUP BY - SQL vs Python Comparison



```
SELECT sex, count(*)
FROM tips
GROUP BY sex;
/*
Female      87
Male       157
*/
```

The pandas equivalent would be:



```
In [17]: tips.groupby('sex').size()
Out[17]:
sex
Female      87
Male       157
dtype: int64
```

source: pandas website

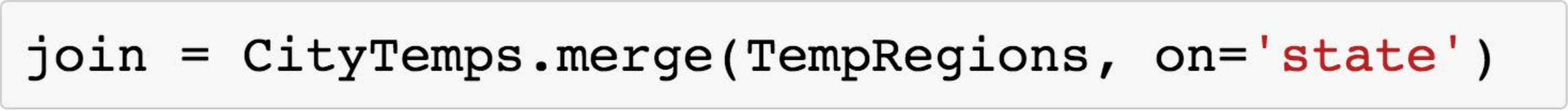
JOINING - SQL vs Python Comparison

SQL QUERY

Select *

From CityTemps, TempsRegions

Where CityTemps.state = TempsRegions.state



```
join = CityTemps.merge(TempRegions, on='state')
```

Useful Links

- 10 minute intro to Pandas
  - http://pandas.pydata.org/pandas-docs/stable/10min.html#min
- Pandas vs SQL comparison
  - http://pandas.pydata.org/pandas-docs/stable/comparison_with_sql.html
- Cheatsheet
  - https://github.com/pandas-dev/pandas/blob/master/doc/cheatsheet/Pandas_Cheat_Sheet.pdf

In Class Demo

In Class Demo

- Download the ***updated*** zipped file on the course website and startup the notebook “2_Data-Operations”

Correlation and Causation

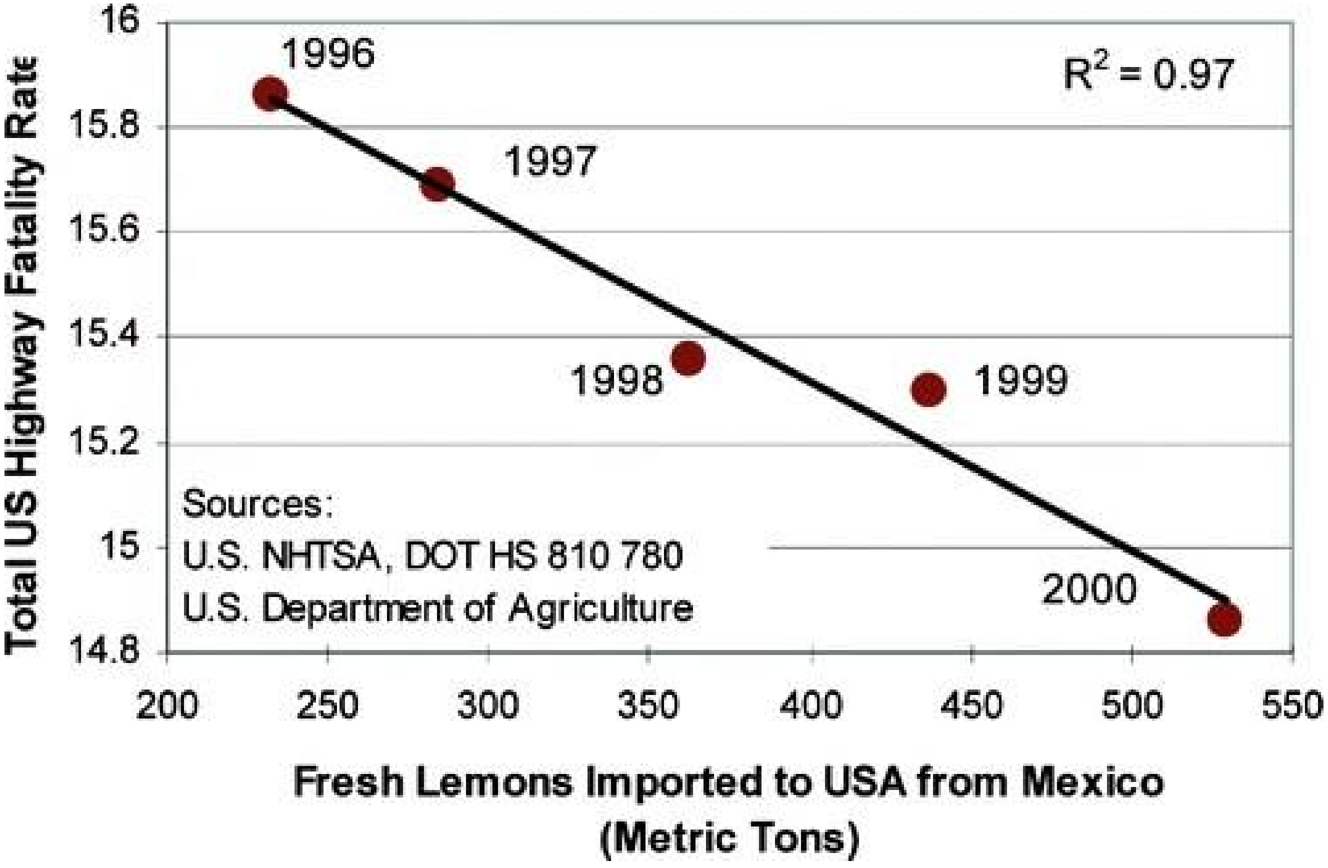
Correlation and Causation

Correlation

Values of **X** and **Y** tend to happen to at the same time

Causation

X’s value tends to influence **Y**’s value



source: http://pubs.acs.org/doi/abs/10.1021/ci700332k

Correlation of chocolate consumption with Nobel Laureates (Image credit: New England Journal of Medicine)

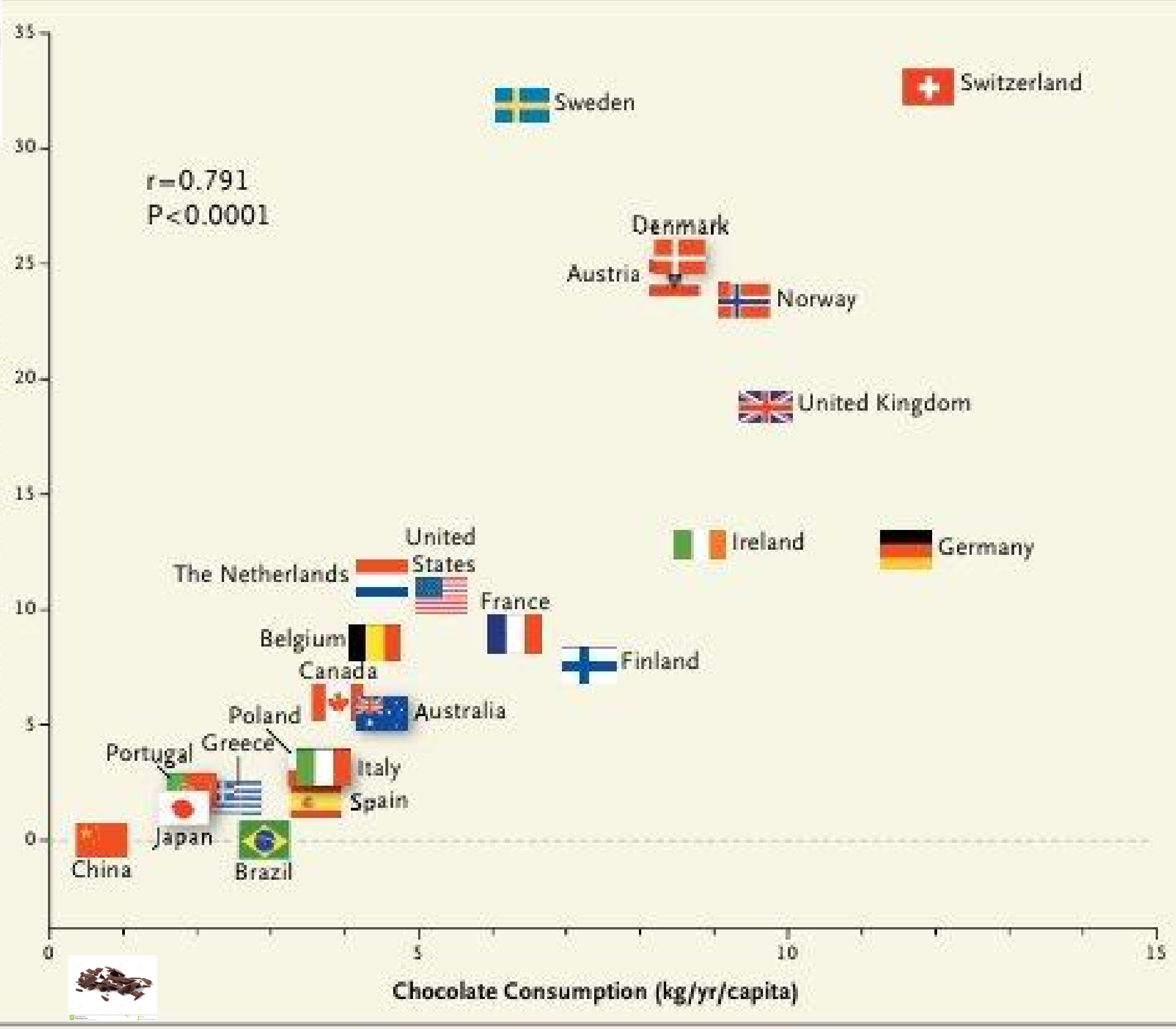


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

An Hour of Running May Add Seven Hours to your life - NYTimes April 12th 2017

Implications

Expected life expectancy in USA

78.74 years

To live till 80..

need to run 1.26 *365 * 24 / 7

= 1576 hours!!!

Exercise pattern

Run one hour a week for 30 years

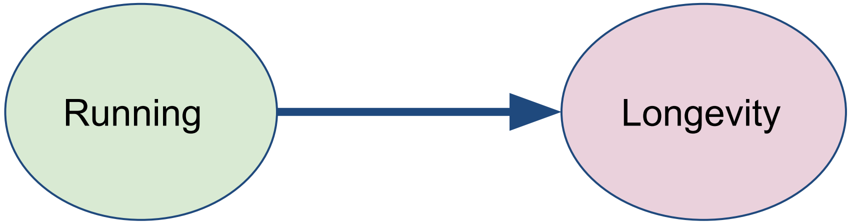
Run 2 hours a week for 15 years

Run 3 hours a week for 10 years

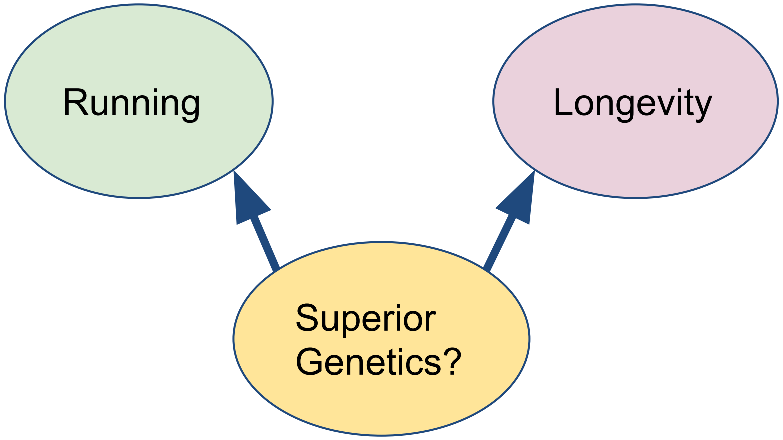
NYTimes article analysis

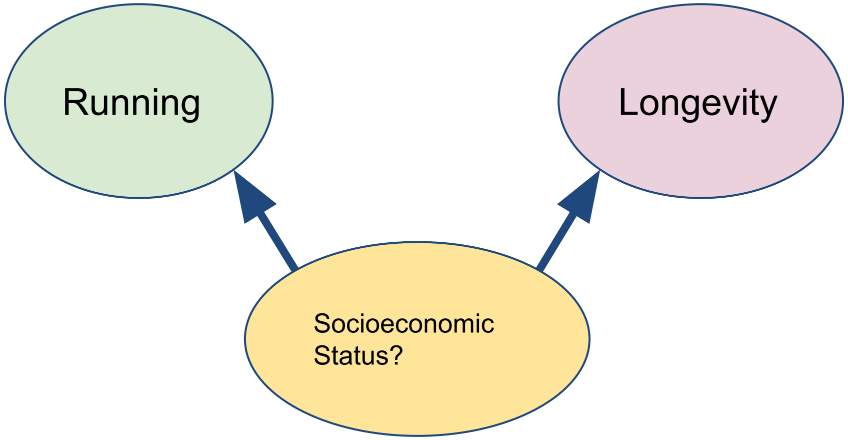
- Title itself already implies some causation

1 hour of running -> 7 hours additional life

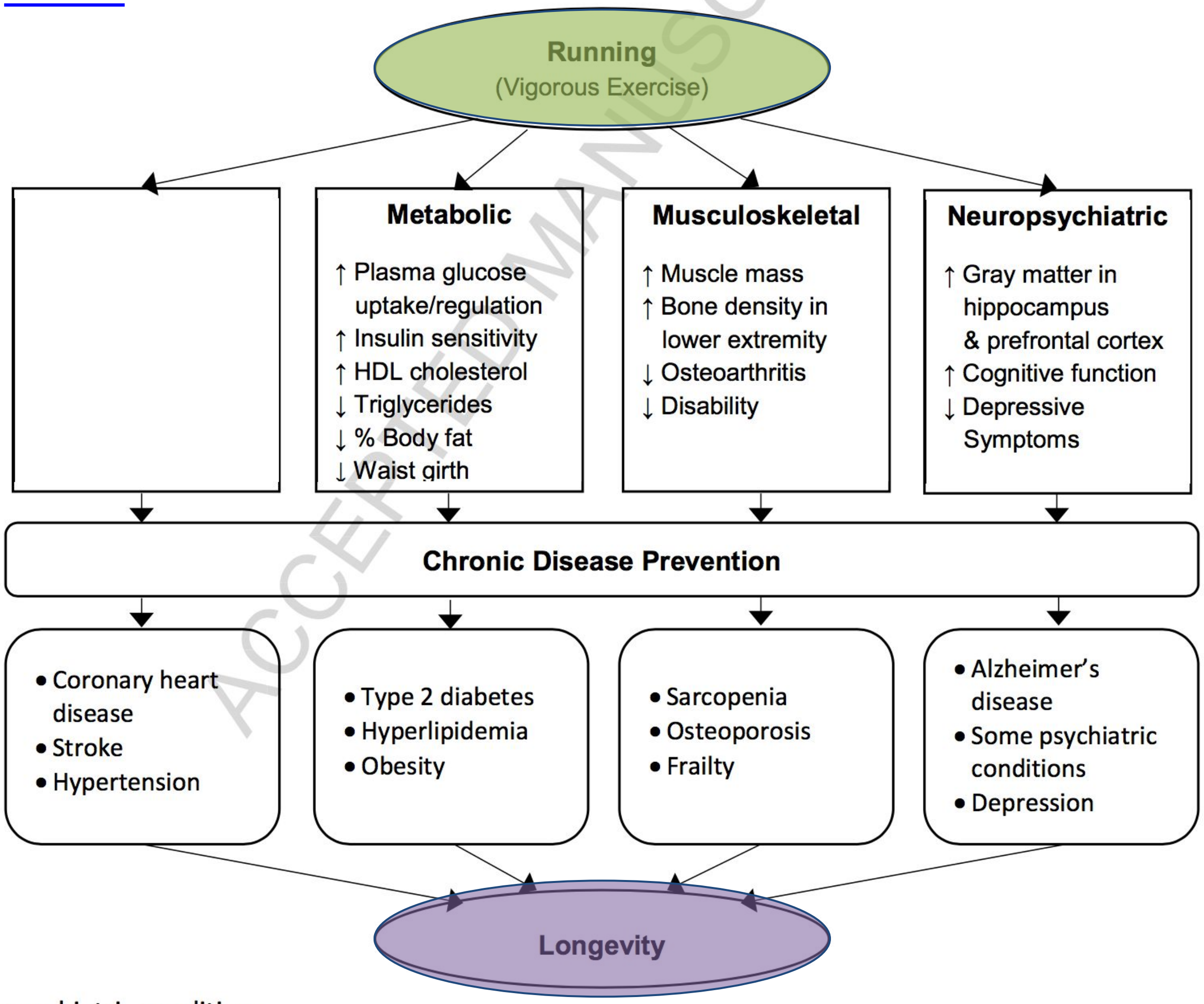


- Possible confounding variables?





Actual study cited



ACCEPTED MANUSCRIPT

Running
(Vigorous Exercise)

Metabolic

- ↑ Plasma glucose uptake/regulation
- ↑ Insulin sensitivity
- ↑ HDL cholesterol
- ↓ Triglycerides
- ↓ % Body fat
- ↓ Waist girth

Musculoskeletal

- ↑ Muscle mass
- ↑ Bone density in lower extremity
- ↓ Osteoarthritis
- ↓ Disability

Neuropsychiatric

- ↑ Gray matter in hippocampus & prefrontal cortex
- ↑ Cognitive function
- ↓ Depressive Symptoms

Chronic Disease Prevention

- Coronary heart disease
- Stroke
- Hypertension

- Type 2 diabetes
- Hyperlipidemia
- Obesity

- Sarcopenia
- Osteoporosis
- Frailty

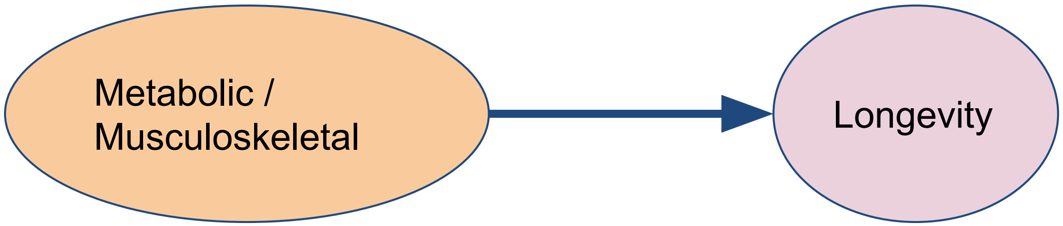
- Alzheimer's disease
- Some psychiatric conditions
- Depression

Longevity

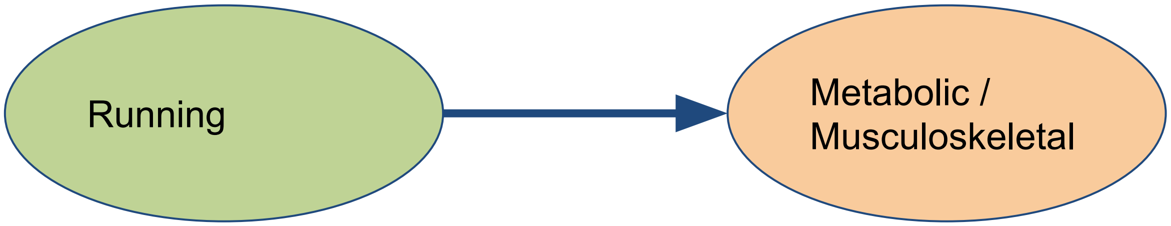
psychiatric conditions.

Figure 3 – Potential Mechanistic Pathway Between Running/Vigorous Exercise and Increased Longevity. An up arrow (↑) indicates an increase and a down arrow (↓) indicates a decrease.

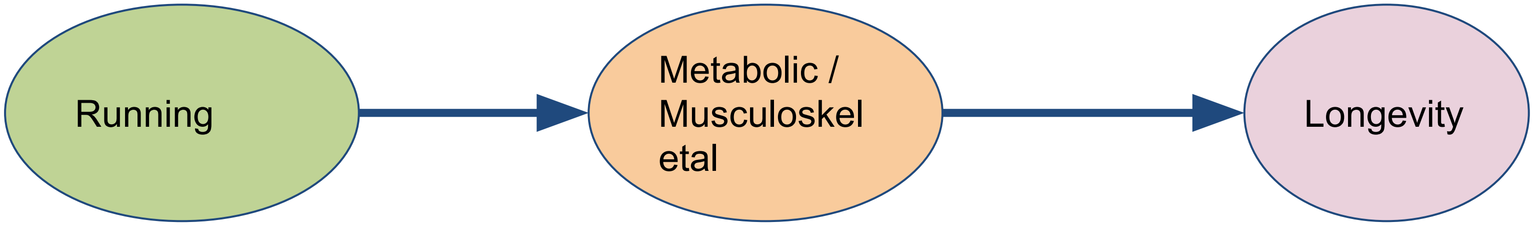
Some previous study concluded that some health statistics are correlated with longer life spans



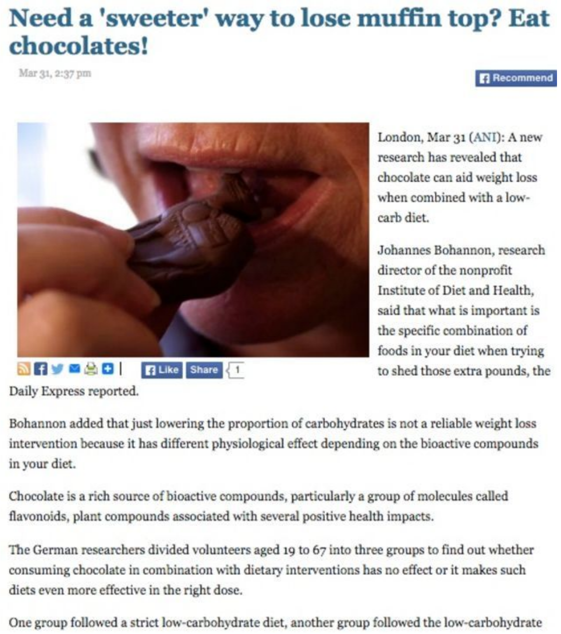
New study determines that running improves a bunch of body health statistics



Therefore, they conclude that running improves longevity



I Fooled Millions Into Thinking Chocolate Helps Weight Loss. Here's How.



Need a 'sweeter' way to lose muffin top? Eat chocolates!

Mar 31, 2:37 pm

London, Mar 31 (ANI): A new research has revealed that chocolate can aid weight loss when combined with a low-carb diet.

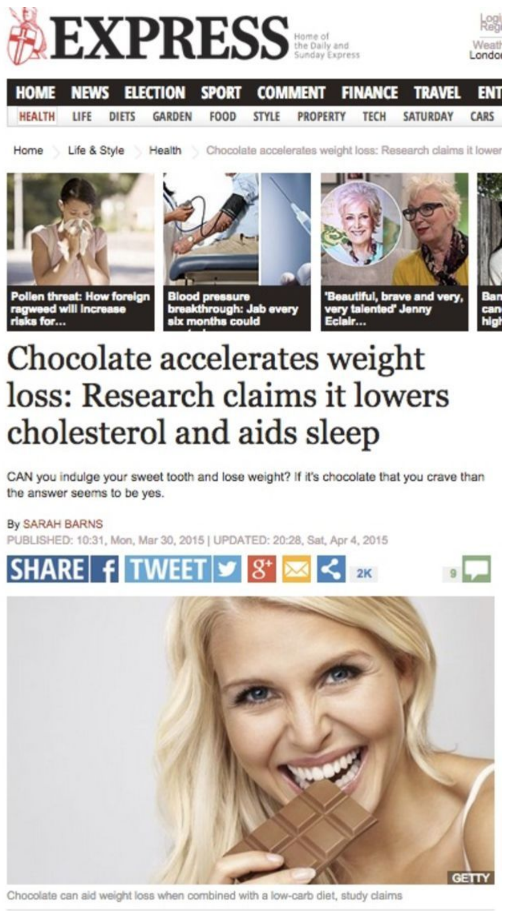
Johannes Bohannon, research director of the nonprofit Institute of Diet and Health, said that what is important is the specific combination of foods in your diet when trying to shed those extra pounds, the Daily Express reported.

Bohannon added that just lowering the proportion of carbohydrates is not a reliable weight loss intervention because it has different physiological effect depending on the bioactive compounds in your diet.

Chocolate is a rich source of bioactive compounds, particularly a group of molecules called flavonoids, plant compounds associated with several positive health impacts.

The German researchers divided volunteers aged 19 to 67 into three groups to find out whether consuming chocolate in combination with dietary interventions has no effect or it makes such diets even more effective in the right dose.

One group followed a strict low-carbohydrate diet, another group followed the low-carbohydrate



EXPRESS
Home of the Daily and Sunday Express
HOME NEWS ELECTION SPORT COMMENT FINANCE TRAVEL ENT
HEALTH LIFE DIETS GARDEN FOOD STYLE PROPERTY TECH SATURDAY CARS
Home
Life & Style
Health
Chocolate accelerates weight loss: Research claims it lower
Pollen threat: How foreign ragweed will increase risks for...
Blood pressure breakthrough: Jab every six months could
'Beautiful, brave and very, very talented' Jenny Eclair...

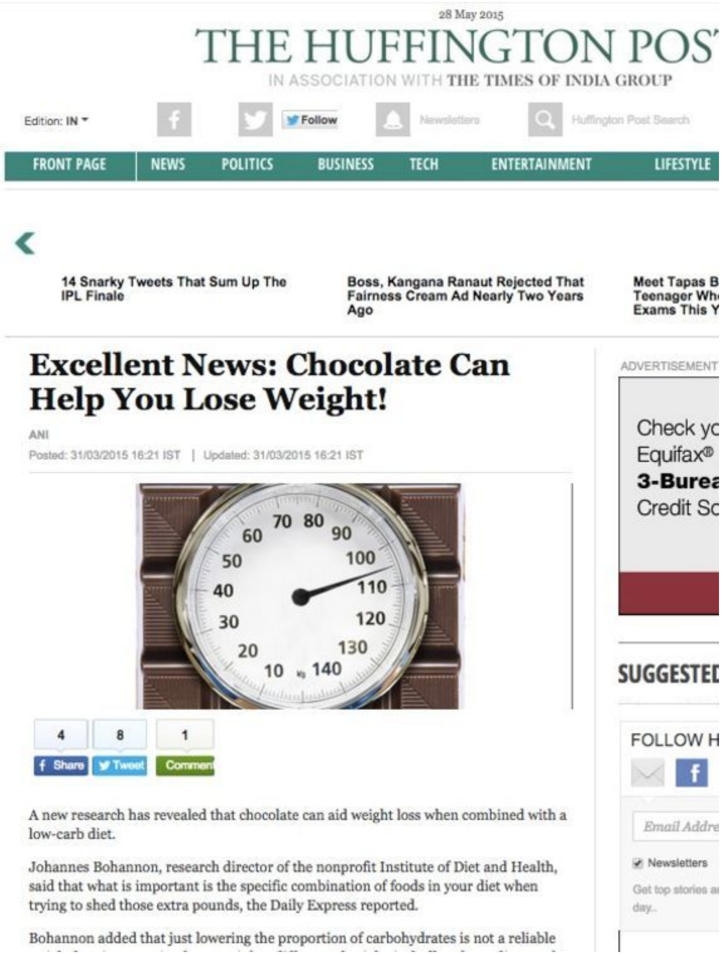
Chocolate accelerates weight loss: Research claims it lowers cholesterol and aids sleep

CAN you indulge your sweet tooth and lose weight? If it's chocolate that you crave than the answer seems to be yes.

By SARAH BARNS

PUBLISHED: 10:31, Mon, Mar 30, 2015 | UPDATED: 20:28, Sat, Apr 4, 2015

Chocolate can aid weight loss when combined with a low-carb diet, study claims



28 May 2015

THE HUFFINGTON POS
IN ASSOCIATION WITH THE TIMES OF INDIA GROUP
Edition: IN
Follow
Newsletters
Huffington Post Search
FRONT PAGE NEWS POLITICS BUSINESS TECH ENTERTAINMENT LIFESTYLE
14 Snarky Tweets That Sum Up The IPL Finale
Boss, Kangana Ranaut Rejected That Fairness Cream Ad Nearly Two Years Ago
Meet Tapas B Teenager Wh Exams This Y
ADVERTISEMENT
Check yo Equifax® 3-Burea Credit Sc
SUGGESTED
FOLLOW H
Email Addre
Newsletters
Get top stories a day..

Excellent News: Chocolate Can Help You Lose Weight!

ANI

Posted: 31/03/2015 16:21 IST | Updated: 31/03/2015 16:21 IST

60
70
80
90
50
100
40
110
30
120
20
130
10
kg
140
4
8
1
Share
Tweet
Comment

A new research has revealed that chocolate can aid weight loss when combined with a low-carb diet.

Johannes Bohannon, research director of the nonprofit Institute of Diet and Health, said that what is important is the specific combination of foods in your diet when trying to shed those extra pounds, the Daily Express reported.

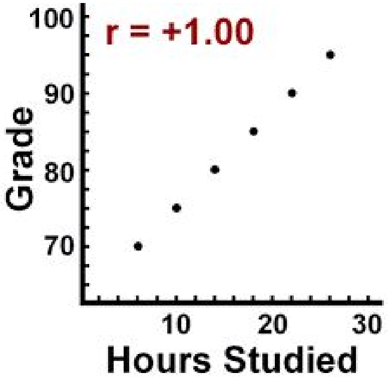
Bohannon added that just lowering the proportion of carbohydrates is not a reliable

Correlation

Positive Correlation

X increases, Y Increases

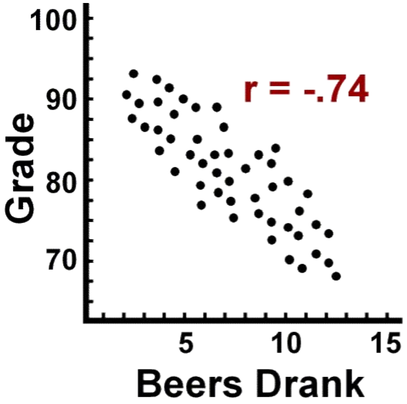
Y increases, X increases



Negative Correlation

X increases, Y decreases

Y increases, X decreases



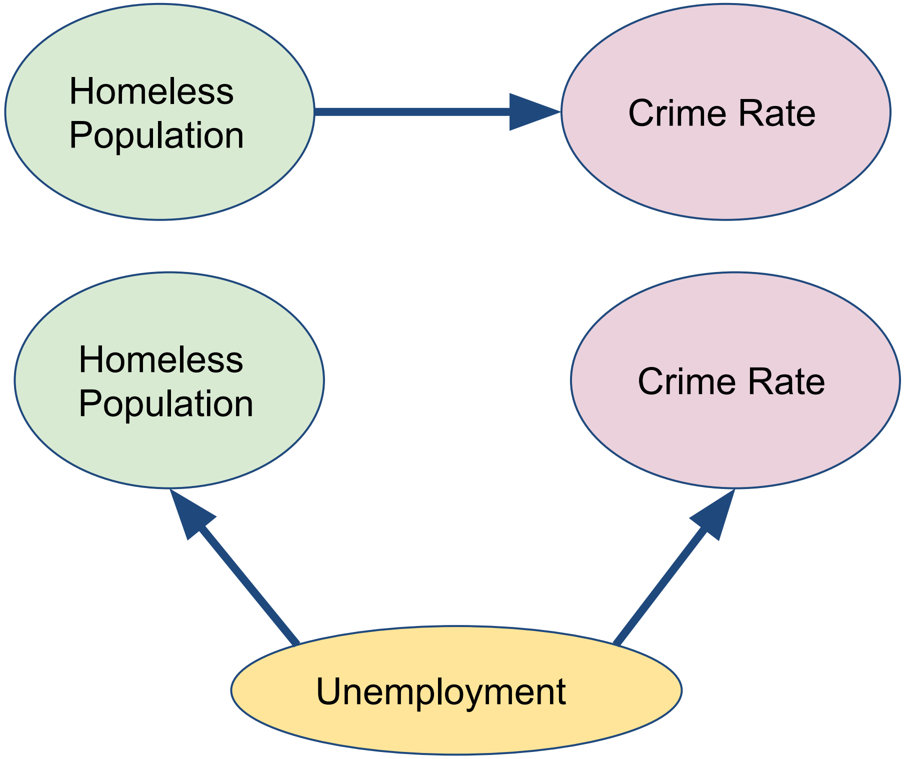
http://www.oxford-review.com/understanding-research-the-direction-of-a-correlation/

Find one example where correlation does not imply causation.

(dicuss in groups of 2-3)

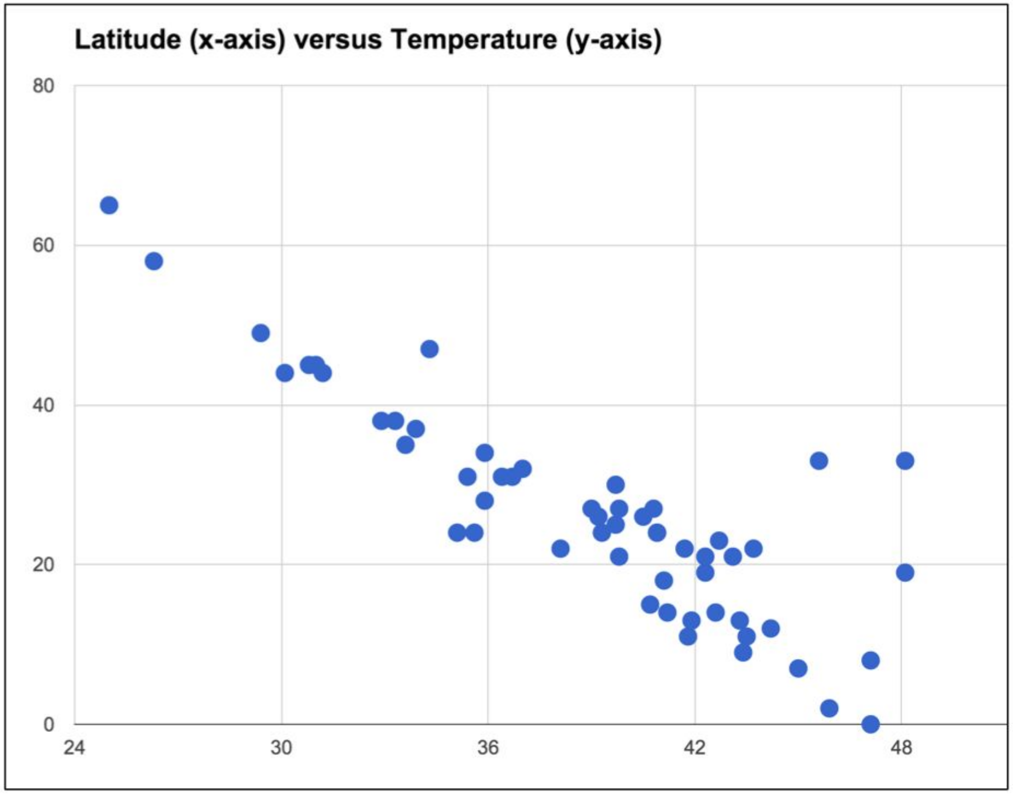
Confounding Variables

Correlation can be result of causation from a hidden **confounding variable**



Determining Correlation

X and Y both ordered: scatterplot

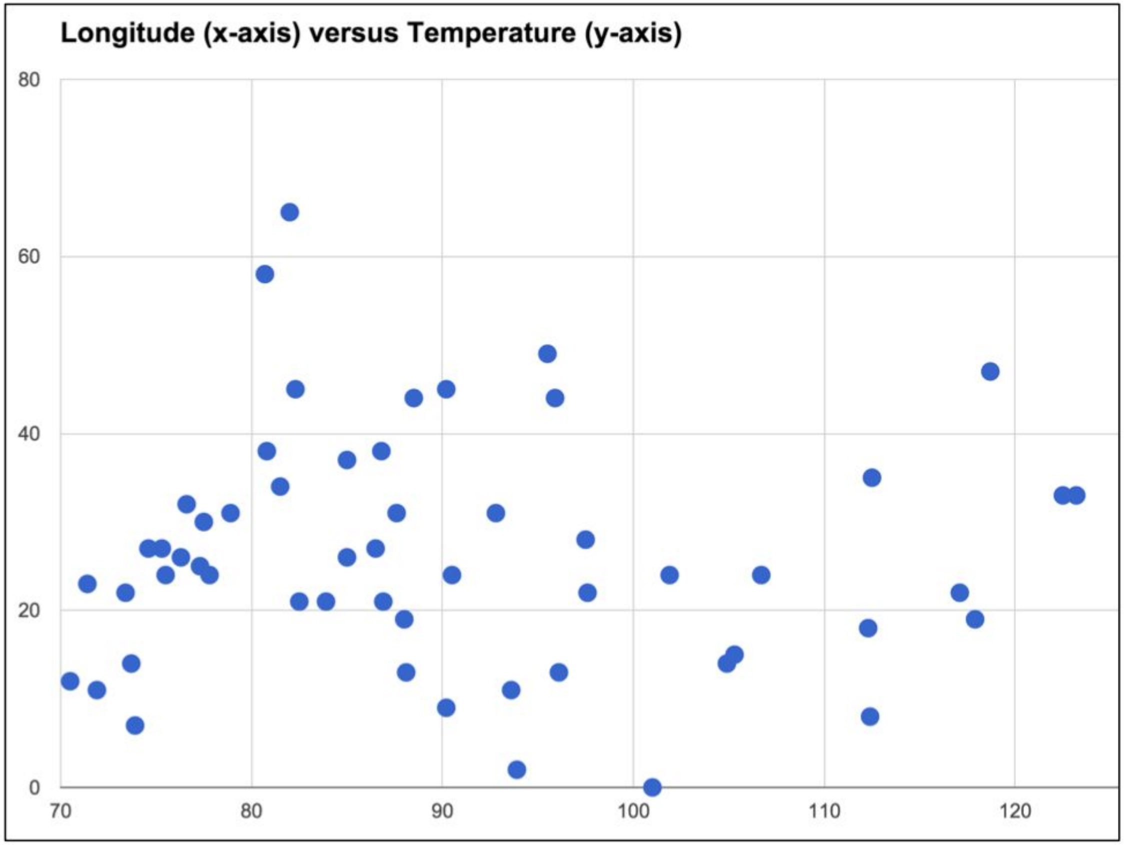


Note: Will cover strength of correlation in ML Regression Lecture

source: Professor Widom's slides

Determining Correlation

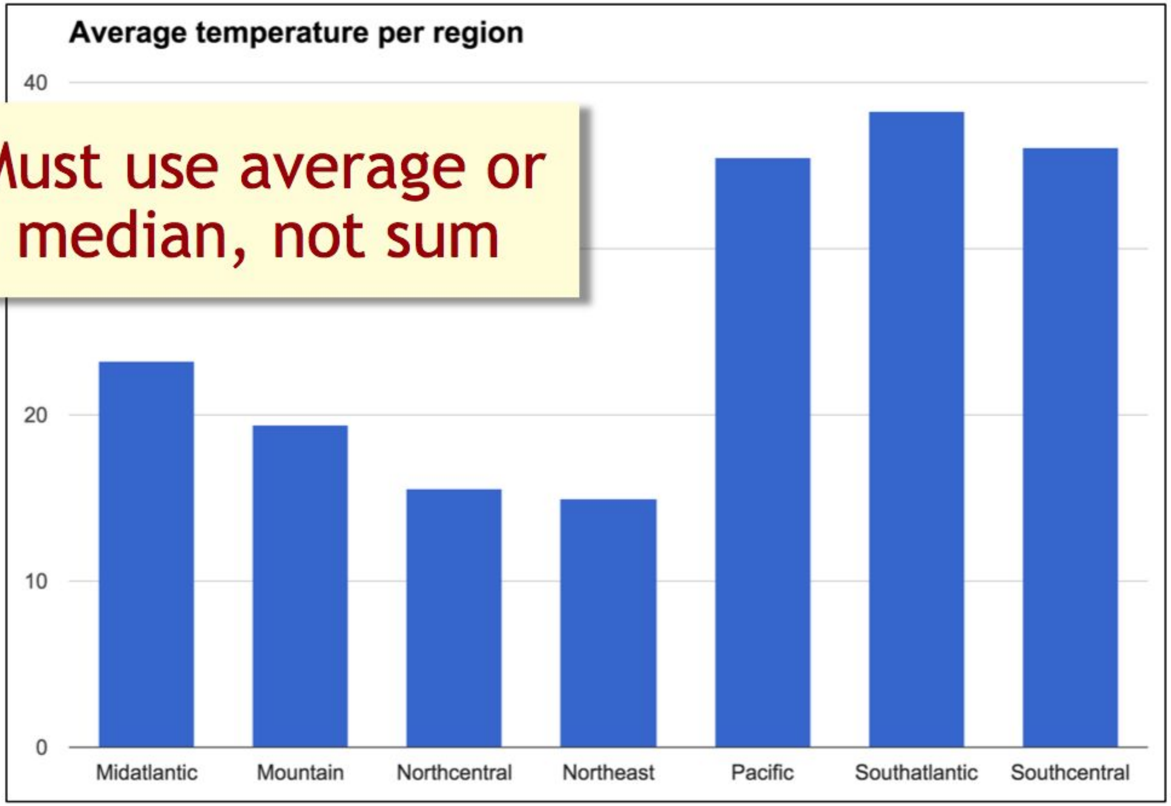
X and Y both ordered: scatterplot



source: Professor Widom’s slides

Determining Correlation

X categorical, Y ordered: bar graph

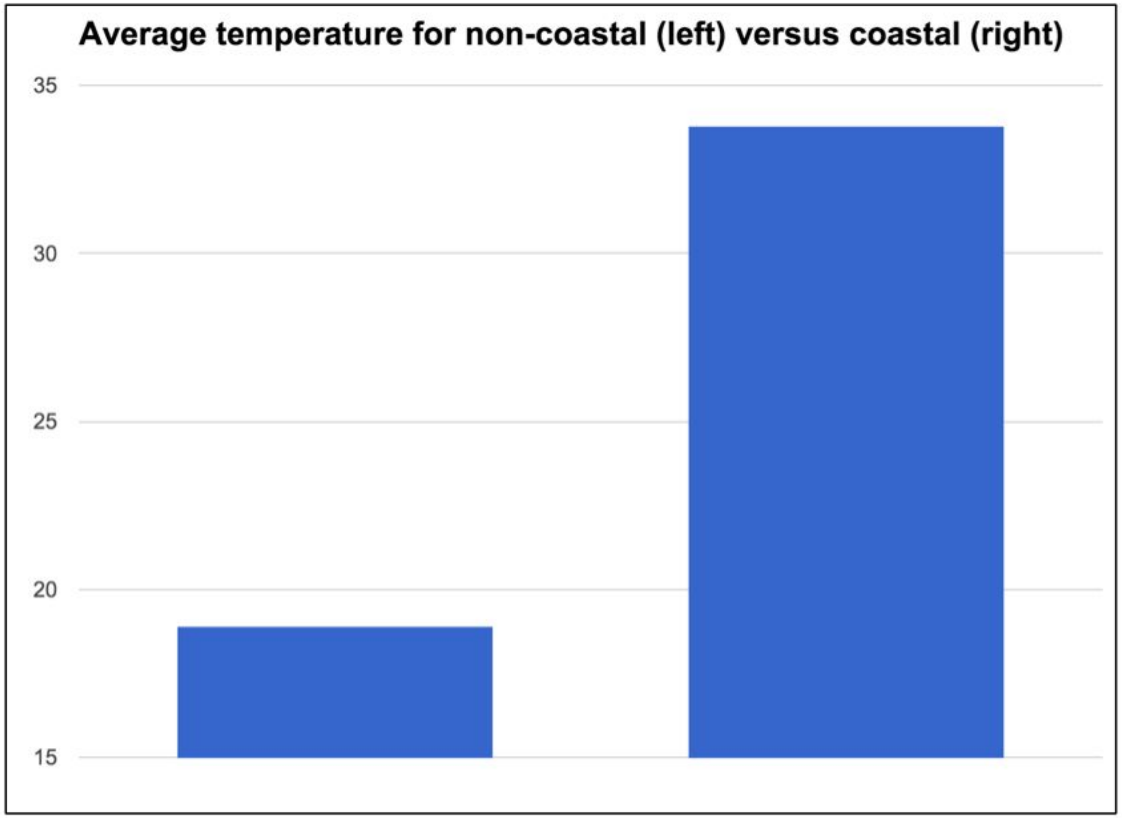


Must use average or median, not sum

source: Professor Widom’s slides

Determining Correlation

X categorical, Y ordered: bar graph



source: Professor Widom's slides

Determining Causation

- Hill’s Criteria
- Running experiments and testing statistical significance

Hill’s Criteria

(Useful guidelines for investigating causality in epidemiological studies)

Strength - of correlation between X and Y

Consistency - of correlation across different datasets

Specificity - no other likely explanation for correlation

Temporality - Y occurs after X

Plausibility - there’s a reason for causation

Coherence - consistent with related theories

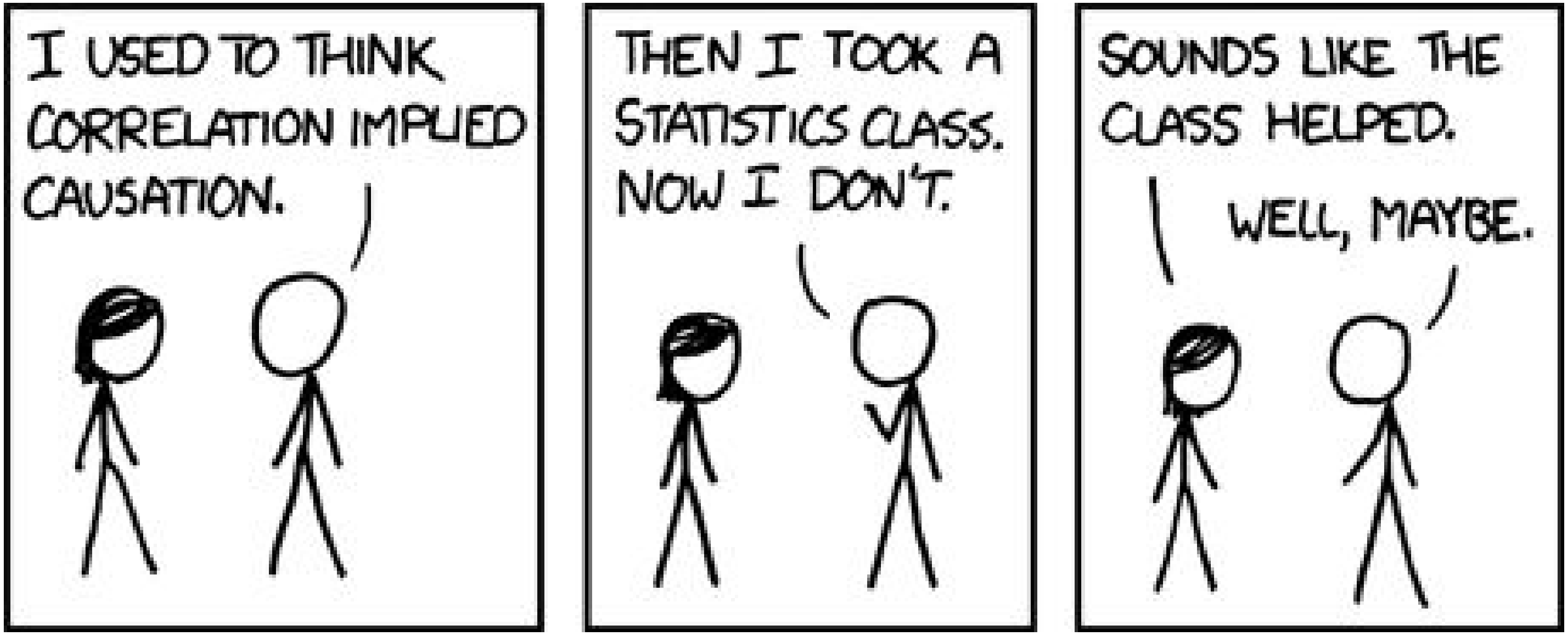
Experimental Validation

Determine if X causes Y

a. Experimental group
b. Control group
c. Ensure no distinctions between 2 groups
   i. If there is, more complicated measures work
d. Determine if there is a statistical significant difference between experimental and control (*not the focus of this class*)

SOMEONE SENT ME ANOTHER ANONYMOUS EMAIL WITH A LINK TO AN ARTICLE ABOUT THE WORLD'S WORST BOSSES.
I GET ONE OF THOSE EMAILS EVERY TIME I LEAVE YOUR CUBICLE. DID YOU THINK I WOULDN'T NOTICE THE CORRELATION?
CORRELATION DOES NOT IMPLY CAUSATION.
SEND
Dilbert.com DilbertCartoonist@gmail.com
11-28-11 ©2011 Scott Adams, Inc./Dist. by Universal Uclick

Correlation vs Causation


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Privacy

Privacy

- Individual data collected covertly
- Data collected legally, used questionably
- Individiuals identified from “anonymous” data
- Invisible Opt in

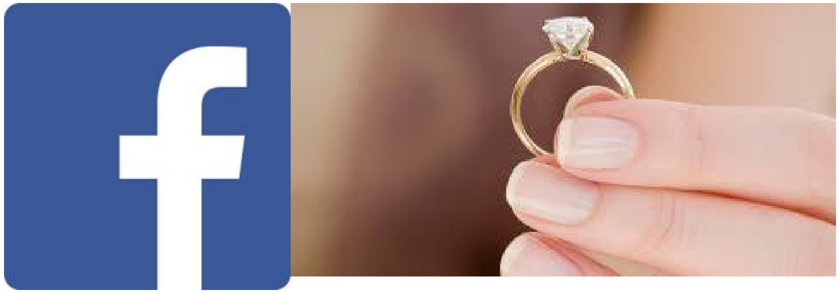
Individual Data Collected Covertly

- National Security Agency collecting phone records and internet traffic of most Americans
  - Revealed by whistleblower Edward Snowden
  - NSA argued “metadata” does not invade privacy
But easy to detect medical conditions, psychological conditions, criminal activity
  - Since ruled illegal

Data Collected Legally, Used Unethically?

- Target Figuring out teen girl was pregnant

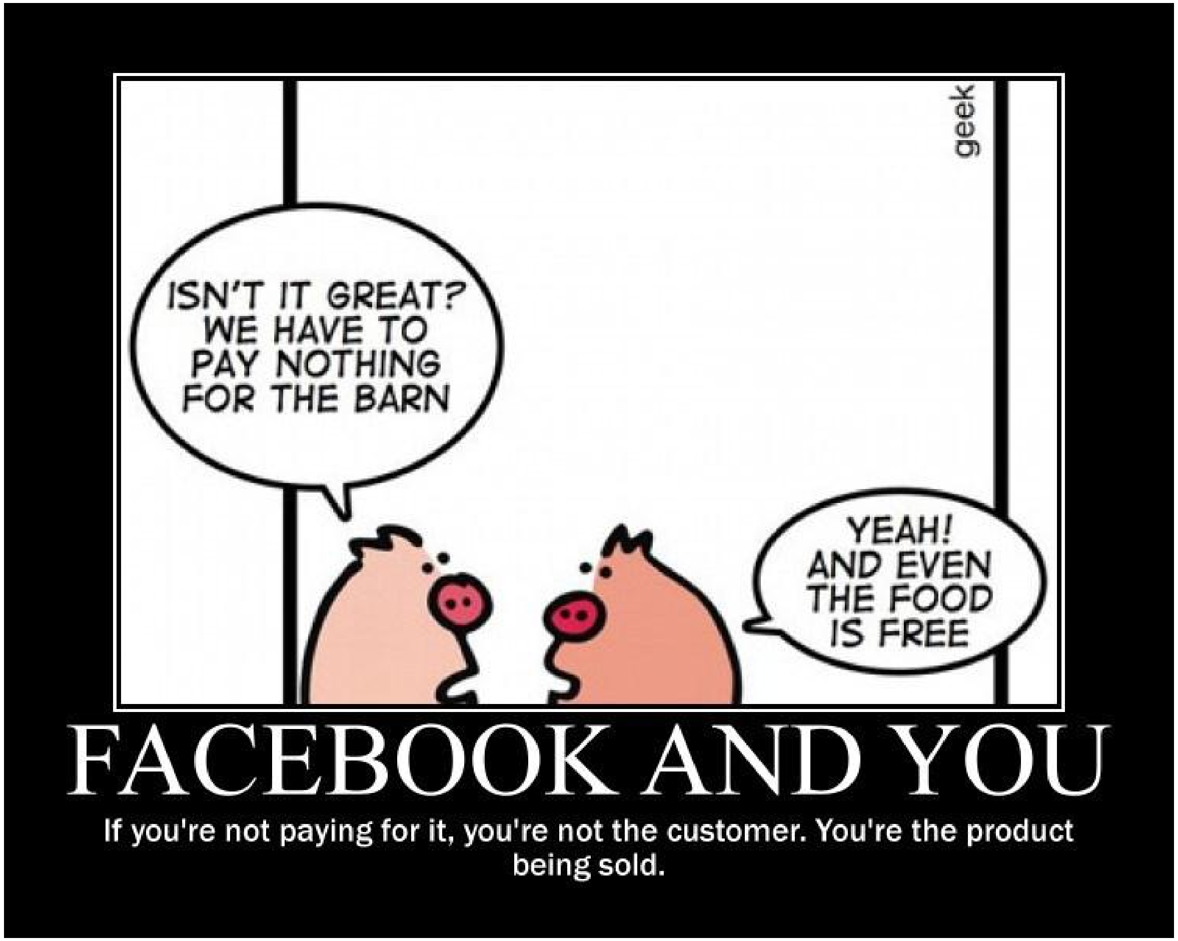
- Facebook “Beacon” Diamond Engagement Ring
  - Facebook publicly posted that a person just bought a diamond ring, ruining the surprise



Deducing Individual Data

- Boston Mayor Health Record’s Identified
    - Easy to identify a person 5-digit zip + gender + date of birth uniquely identifies 87% of U.S. population
- Netflix users identified
    - Researchers were able to identify targets matching netflix reviews with data from IMDb.
    - If you knew a few movies a subscriber rented in a time period

Invisible Opt In

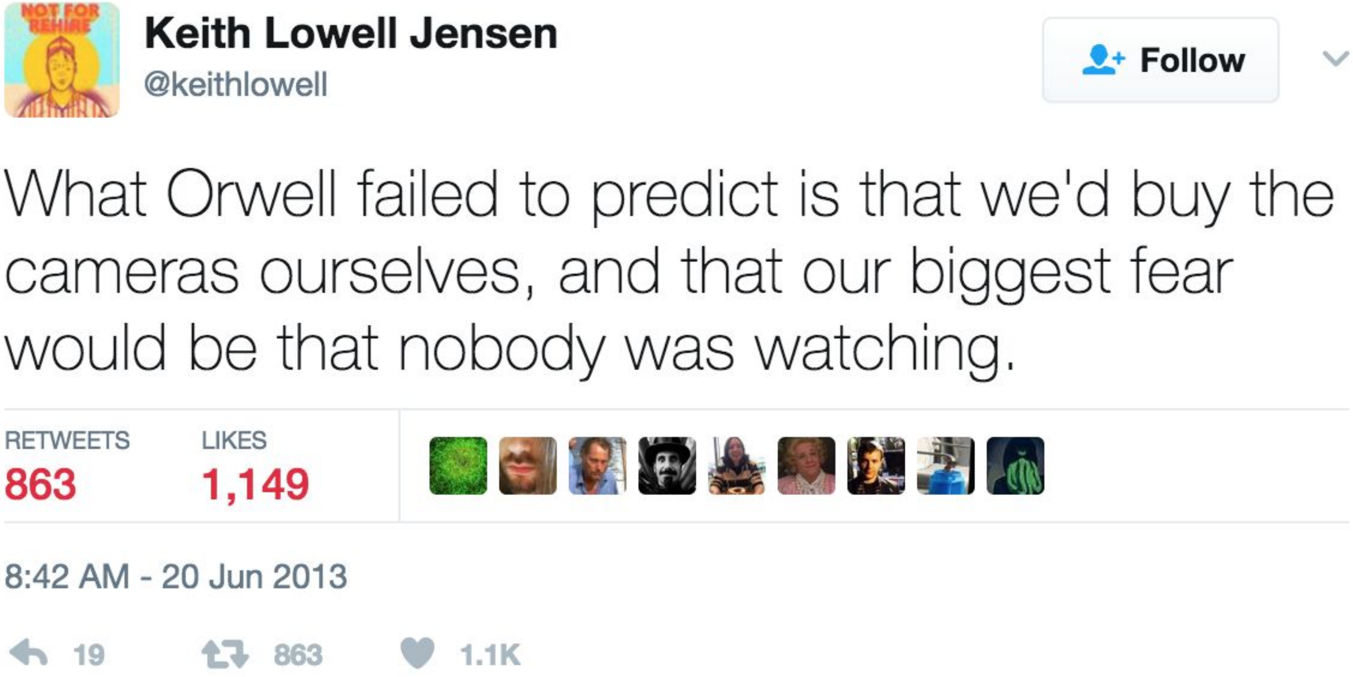

geek
ISN'T IT GREAT? WE HAVE TO PAY NOTHING FOR THE BARN
YEAH! AND EVEN THE FOOD IS FREE
FACEBOOK AND YOU
If you're not paying for it, you're not the customer. You're the product being sold.

Invisible Opt In

Cookies are small files stored on a user's computer to hold data specific to a user and website

Can track you across the web, you have cookies turned by default

1984 Big Brother? Or...



Keith Lowell Jensen
@keithlowell

What Orwell failed to predict is that we'd buy the cameras ourselves, and that our biggest fear would be that nobody was watching.

RETWEETS 863
LIKES 1,149

8:42 AM - 20 Jun 2013

End