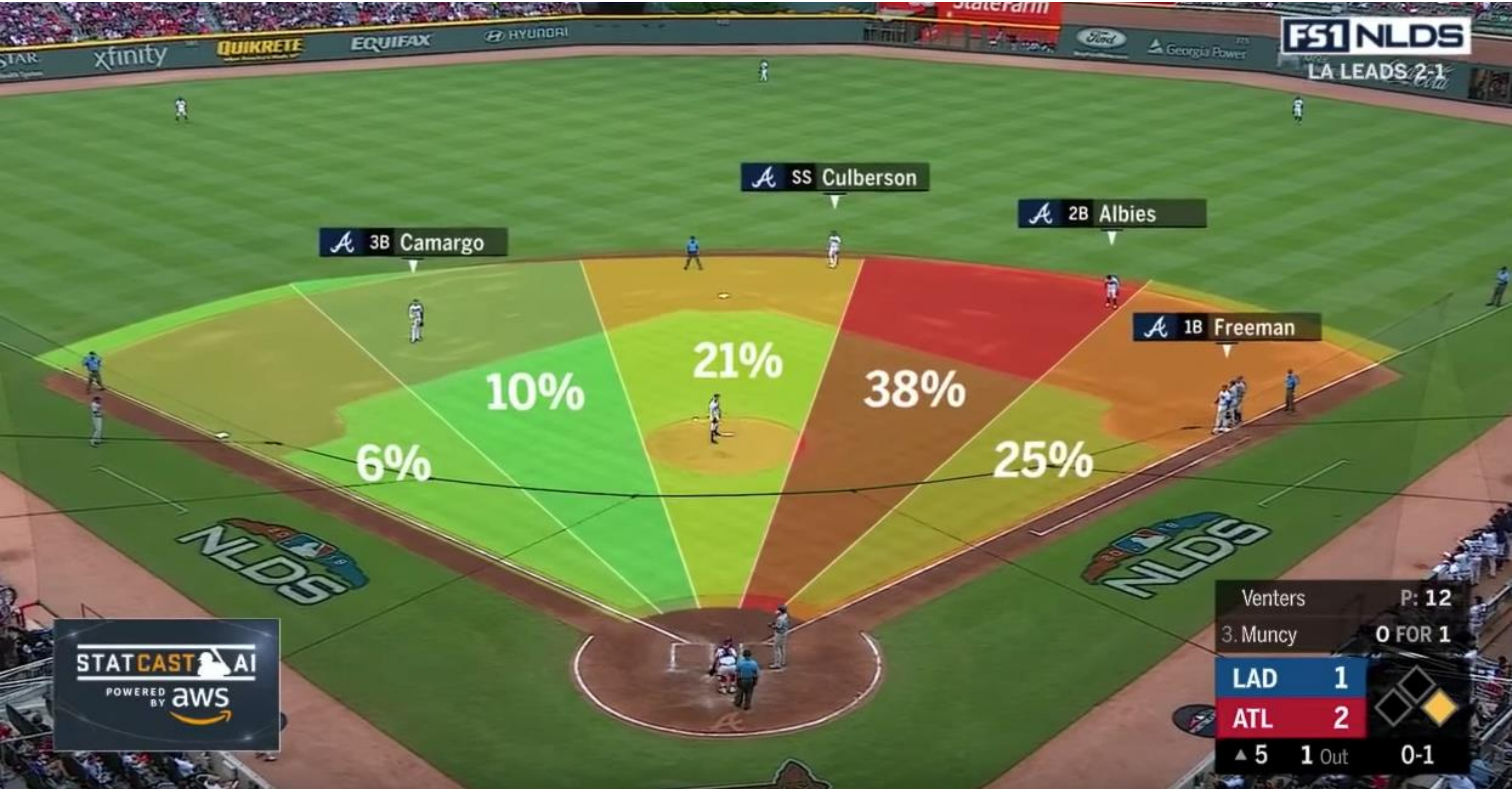
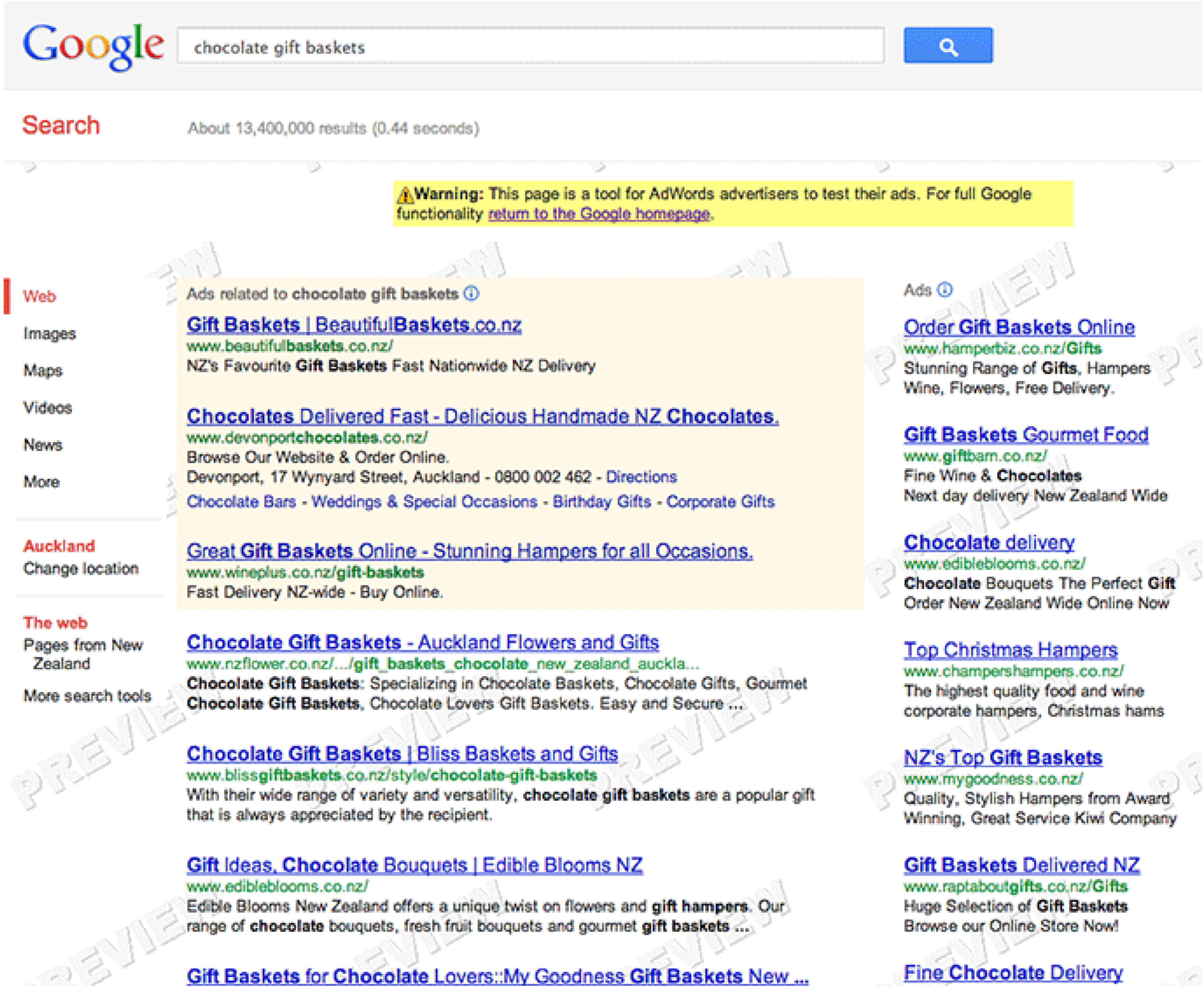
CSCE 689: Special Topics in Modern Algorithms for Data Science

Lecture 1

Samson Zhou

Why Data Science?


FS1 NLDS
LA LEADS 2-1
xfinity
QUIKRETE
EQUIFAX
HYUNDAI
Ford
Georgia Power
SS Culberson
2B Albies
3B Camargo
1B Freeman
21%
10%
38%
6%
25%
NLDS
NLDS
STATCAST AI
POWERED BY aws
Venters
P: 12
3. Muncy
0 FOR 1
LAD 1
ATL 2
▲5
1 Out
0-1



Search

About 13,400,000 results (0.44 seconds)

⚠**Warning:** This page is a tool for AdWords advertisers to test their ads. For full Google functionality return to the Google homepage.

Ads related to **chocolate gift baskets** ⓘ

Gift Baskets | Beautiful**Baskets**.co.nz
www.beautiful**baskets**.co.nz/
NZ's Favourite **Gift Baskets** Fast Nationwide NZ Delivery

Chocolates Delivered Fast - Delicious Handmade NZ **Chocolates**.
www.devonport**chocolates**.co.nz/
Browse Our Website & Order Online.
Devonport, 17 Wynyard Street, Auckland - 0800 002 462 - Directions
Chocolate Bars - Weddings & Special Occasions - Birthday Gifts - Corporate Gifts

Great **Gift Baskets** Online - Stunning Hampers for all Occasions.
www.wineplus.co.nz/**gift-baskets**
Fast Delivery NZ-wide - Buy Online.

Chocolate Gift Baskets - Auckland Flowers and Gifts
www.nzflower.co.nz/.../**gift_baskets_chocolate**_new_zealand_auckla...
Chocolate Gift Baskets: Specializing in Chocolate Baskets, Chocolate Gifts, Gourmet **Chocolate Gift Baskets**, Chocolate Lovers Gift Baskets. Easy and Secure ...

Chocolate Gift Baskets | Bliss Baskets and Gifts
www.bliss**giftbaskets**.co.nz/style/**chocolate-gift-baskets**
With their wide range of variety and versatility, **chocolate gift baskets** are a popular gift that is always appreciated by the recipient.

Gift Ideas, **Chocolate** Bouquets | Edible Blooms NZ
www.edibleblooms.co.nz/
Edible Blooms New Zealand offers a unique twist on flowers and **gift hampers**. Our range of **chocolate** bouquets, fresh fruit bouquets and gourmet **gift baskets** ...

Gift Baskets for **Chocolate** Lovers::My Goodness **Gift Baskets** New ...

Ads ⓘ

Order **Gift Baskets** Online
www.hamperbiz.co.nz/**Gifts**
Stunning Range of **Gifts**, Hampers Wine, Flowers, Free Delivery.

Gift Baskets Gourmet Food
www.**gift**barn.co.nz/
Fine Wine & **Chocolates**
Next day delivery New Zealand Wide

Chocolate delivery
www.edibleblooms.co.nz/
Chocolate Bouquets The Perfect **Gift**
Order New Zealand Wide Online Now

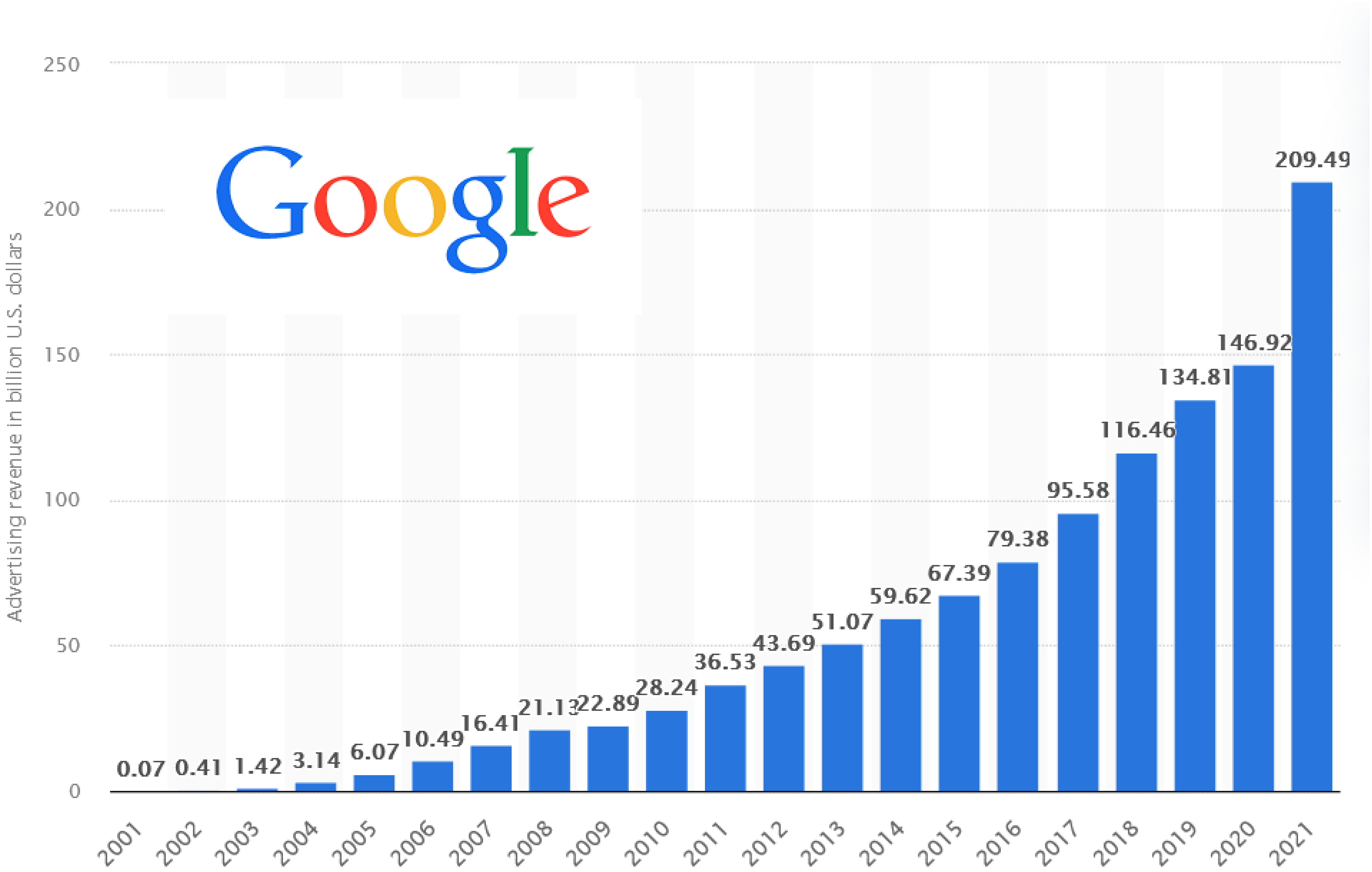
Top Christmas Hampers
www.champershampers.co.nz/
The highest quality food and wine corporate hampers, Christmas hams

NZ's Top **Gift Baskets**
www.mygoodness.co.nz/
Quality, Stylish Hampers from Award Winning, Great Service Kiwi Company

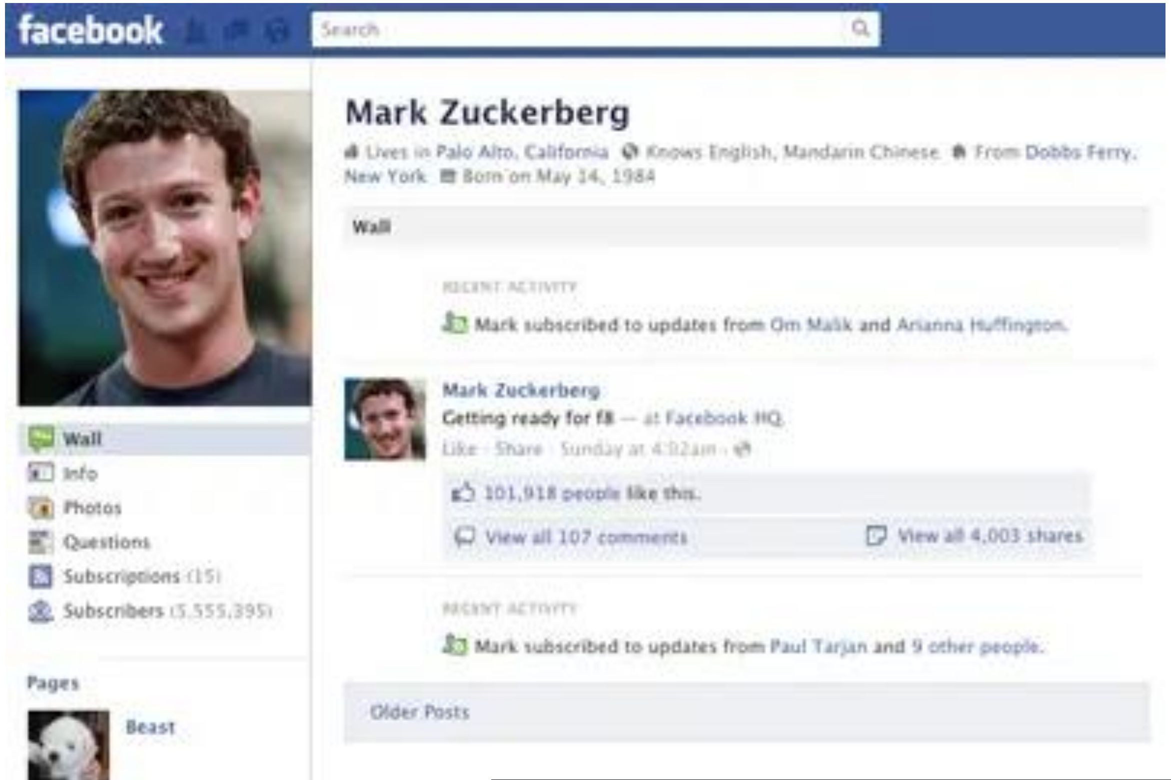
Gift Baskets Delivered NZ
www.raptabout**gifts**.co.nz/**Gifts**
Huge Selection of **Gift Baskets**
Browse our Online Store Now!

Fine **Chocolate** Delivery

Google Ad Revenue (2001-2021, billion USD)



© Statista 2023



Mark Zuckerberg

Lives in Palo Alto, California · Knows English, Mandarin Chinese · From Dobbs Ferry, New York · Born on May 14, 1984

Wall

RECENT ACTIVITY

Mark subscribed to updates from Om Malik and Arianna Huffington.

Mark Zuckerberg

Getting ready for f8 — at Facebook HQ.

Like · Share · Sunday at 4:02am ·

101,918 people like this.

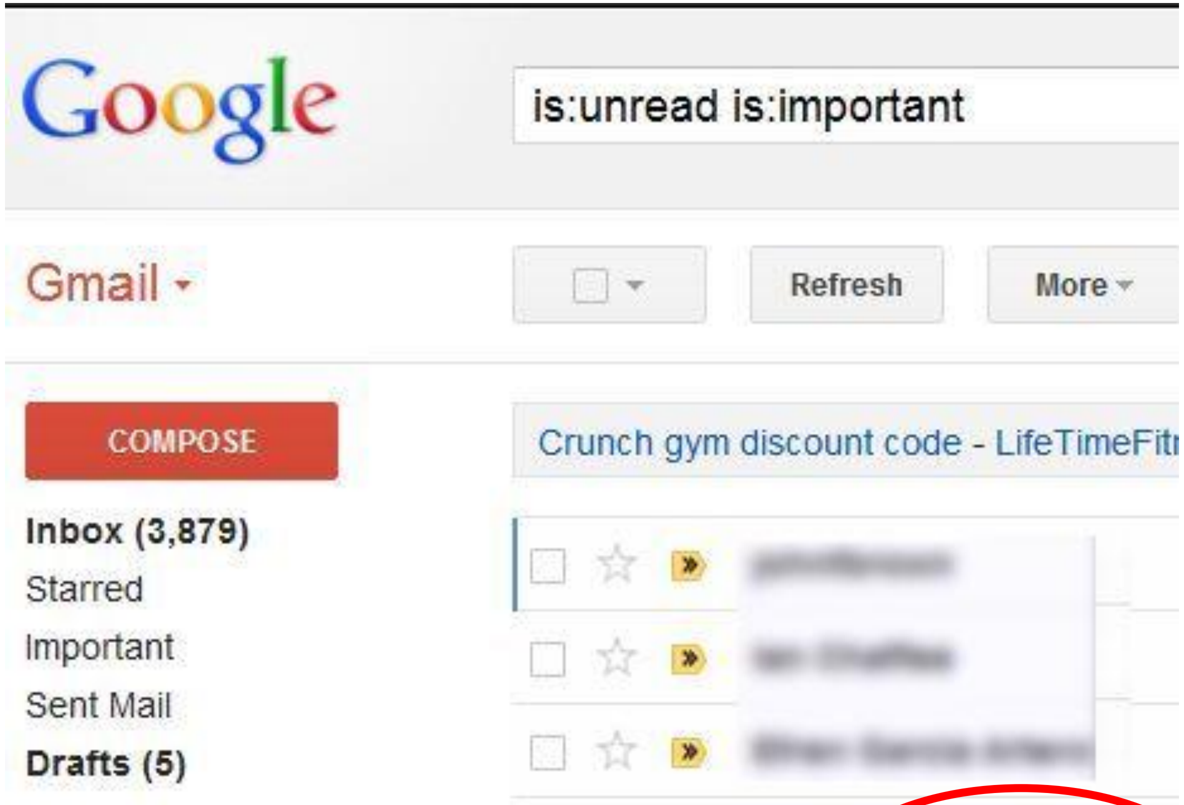
View all 107 comments

View all 4,003 shares

RECENT ACTIVITY

Mark subscribed to updates from Paul Tarjan and 9 other people.

Older Posts



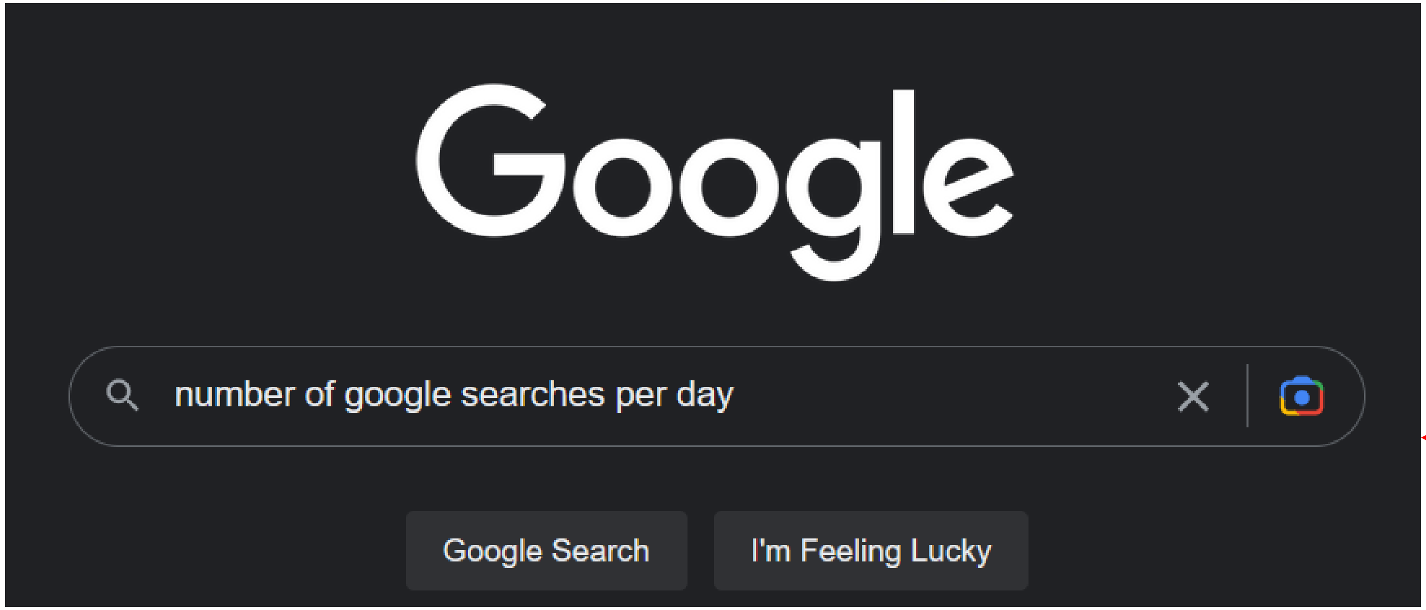
is:unread is:important

Crunch gym discount code - LifeTimeFit

□ ☆ [redacted]

□ ☆ [redacted]

□ ☆ [redacted]



number of google searches per day

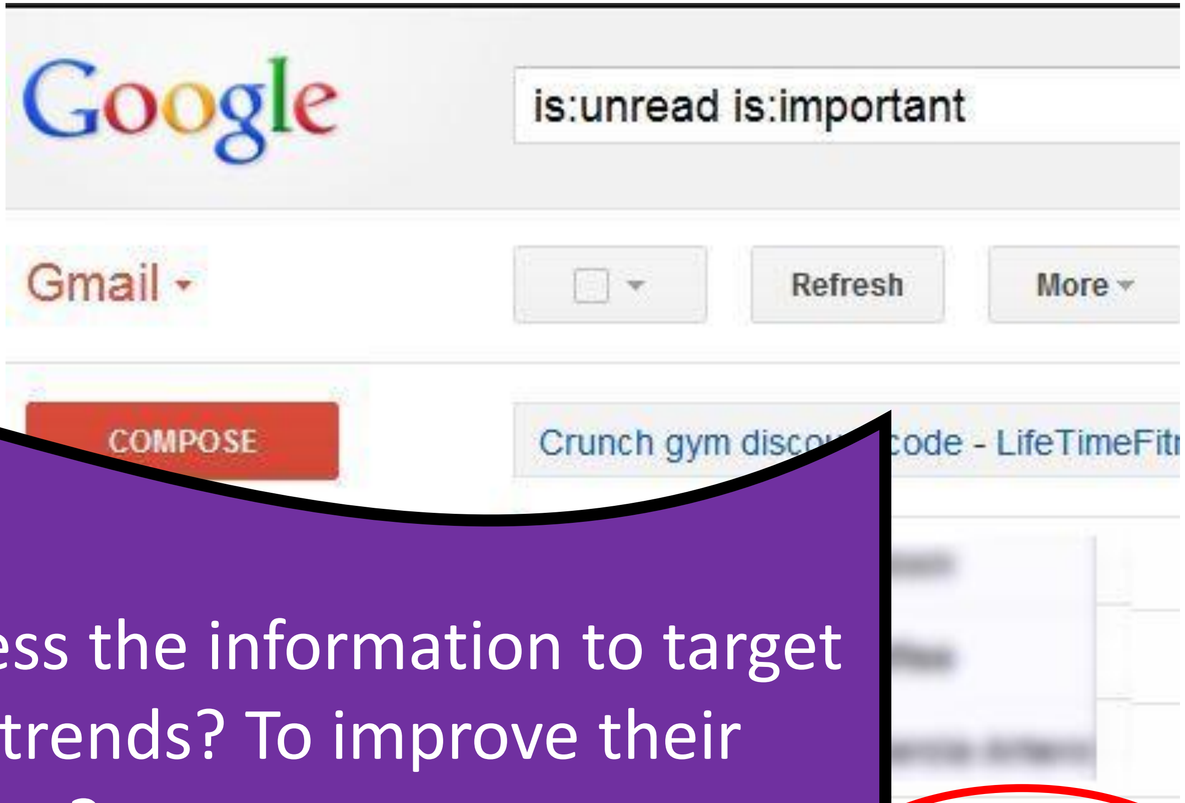
3 billion monthly active users

330 billion daily e-mails

8.5 billion daily Google searches

Mark Zuckerberg

Lives in Palo Alto, California · Knows English, Mandarin Chinese · From Dobbs Ferry, New York · Born on May 14, 1984



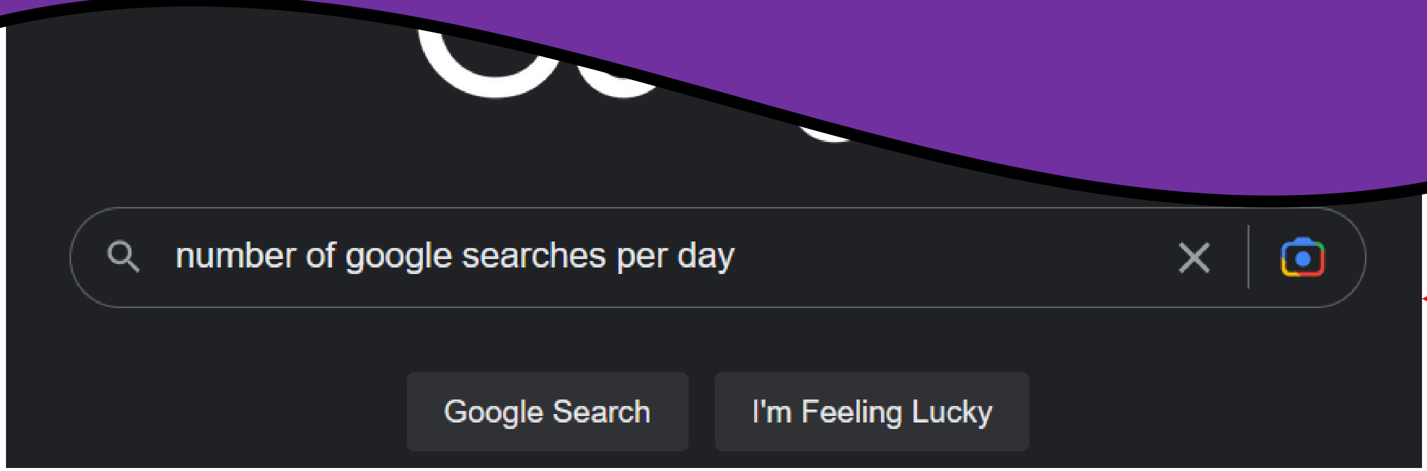
is:unread is:important

Crunch gym disco code - LifeTimeFit

How do these companies process the information to target advertisements? To predict trends? To improve their products?

3 billion
monthly
active users

30 billion
aily e-mails

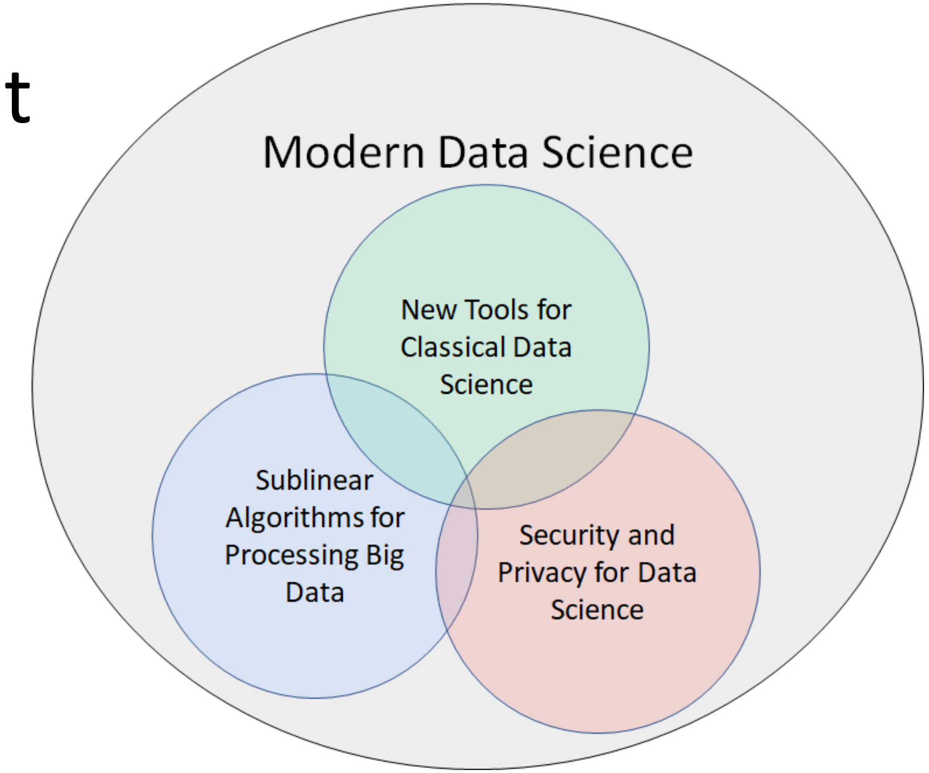


number of google searches per day

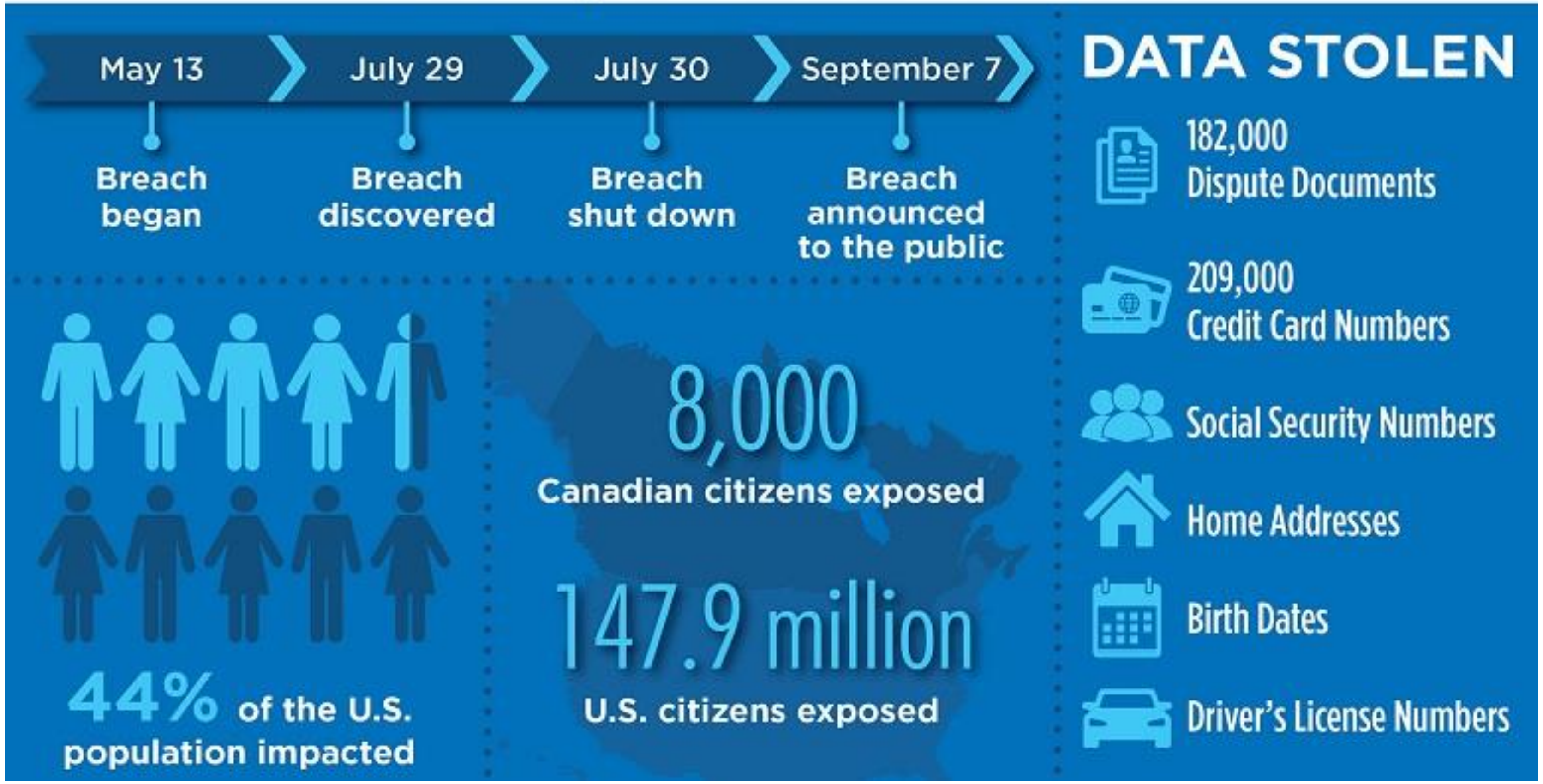
8.5 billion
daily Google
searches

Evolving Demands

- Sublinear-time or sublinear-space algorithms
- Incorporation of advice
- Security and privacy
- Robustness to noise or adversarial input
- Ability to handle time-sensitive data


Modern Data Science
New Tools for Classical Data Science
Sublinear Algorithms for Processing Big Data
Security and Privacy for Data Science

2017 Equifax Data Breach



May 13 — Breach began

July 29 — Breach discovered

July 30 — Breach shut down

September 7 — Breach announced to the public

44% of the U.S. population impacted

8,000 Canadian citizens exposed

147.9 million U.S. citizens exposed

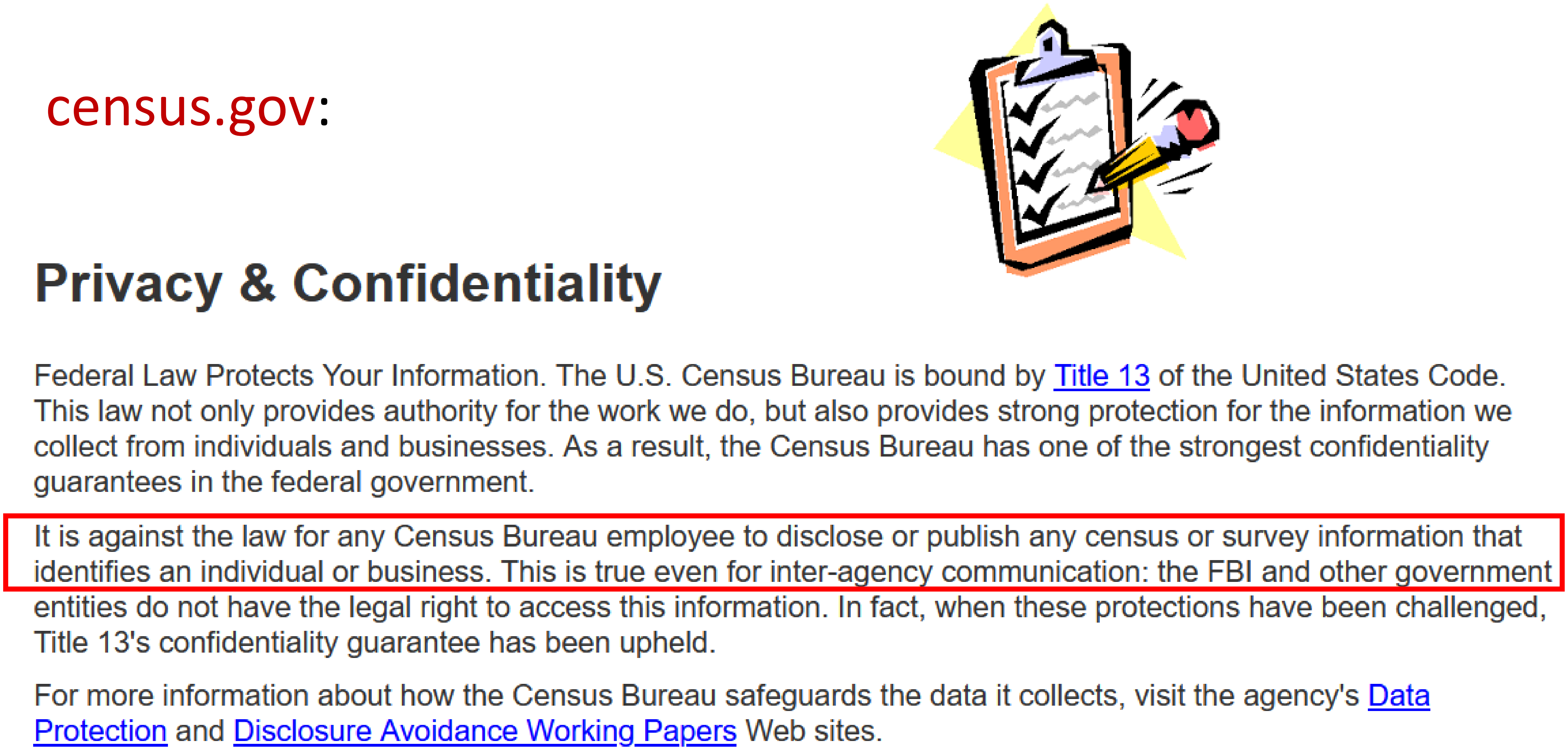
DATA STOLEN

- 182,000 Dispute Documents
- 209,000 Credit Card Numbers
- Social Security Numbers
- Home Addresses
- Birth Dates
- Driver's License Numbers

"Equifax agreed to a $700 million settlement over the privacy breach, but $425 million of that was set aside to repay consumers as a restitution fund."

YAHOO!
UBER
Gmail by Google
Adobe
BLIZZARD ENTERTAINMENT
eHarmony
AMERICA Online
U.S. DEPARTMENT OF HOMELAND SECURITY
LastPass ****
ebay
Dropbox
Linked in
PlayStation

census.gov:



Privacy & Confidentiality

Federal Law Protects Your Information. The U.S. Census Bureau is bound by Title 13 of the United States Code. This law not only provides authority for the work we do, but also provides strong protection for the information we collect from individuals and businesses. As a result, the Census Bureau has one of the strongest confidentiality guarantees in the federal government.

It is against the law for any Census Bureau employee to disclose or publish any census or survey information that identifies an individual or business. This is true even for inter-agency communication: the FBI and other government entities do not have the legal right to access this information. In fact, when these protections have been challenged, Title 13's confidentiality guarantee has been upheld.

For more information about how the Census Bureau safeguards the data it collects, visit the agency's Data Protection and Disclosure Avoidance Working Papers Web sites.

Anonymizing Data

Age	Zip Code	Employer	Has Pet
56	77005	Apple	Yes
32	77005	Microsoft	No
71	77005	Amazon	Yes
44	77005	Petsmart	Yes
25	77005	Netflix	No
61	77005	Google	No

Anonymizing Data

Age	Zip Code	Employer	Has Pet
56	77005	Apple	Yes
32	77005	Microsoft	No
71	77005	Amazon	Yes
44	77005	Petsmart	Yes
25	77005	Netflix	No
61	77005	Google	No

Name	Age	Gender	Employer
Alice	56	Female	Apple
Bob	32	Male	Microsoft
Carol	71	Female	Amazon
Dale	44	Male	Petsmart
Erin	25	Female	Netflix
Fred	61	Male	Google

Reconstruction Attack

Name	Age	Zip Code	Gender	Employer	Has Pet
Alice	56	77005	Female	Apple	Yes
Bob	32	77005	Male	Microsoft	No
Carol	71	77005	Female	Amazon	Yes
Dale	44	77005	Male	Petsmart	Yes
Erin	25	77005	Female	Netflix	No
Fred	61	77005	Male	Google	No

Implications of the simulated attack

The Census Bureau believed in 2010 that it was necessary to coarsen geographic identifiers in microdata such that the minimum population in any published geography was at least 100,000 persons (Public-Use Microdata Areas).

Our simulated reconstruction-abetted re-identification attack demonstrated that the tabular summaries from the 2010 Census can be converted into a 100% microdata file with geographic precision to the census block-level.

Our simulated attack demonstrated that, depending on the quality of the external data used, between 52 and 179 million respondents to the 2010 Census can be correctly re-identified from the reconstructed microdata.

Stronger privacy protections, such as those in the 2020 Census Disclosure Avoidance System, are necessary to protect against reconstruction-abetted attacks.

Class Motivation

- Data Science is highly interdisciplinary and highly evolving

- Many techniques are not covered in traditional CS classes

Modern Algorithms for Data Science

- Algorithms for data science
- Sublinear algorithms
- Models of computation for big data
- Differential privacy

Logistics

- HRBB 126, MWF, 1:50-2:40 pm CT

- Office Hours: PETR 424, 3 pm CT on Wednesdays

- Course materials: https://samsonzhou.github.io/csce689-2023

Primary Goals

- Describe the motivation and statement of central data science problems, measured by the midterm presentation

- Work in various big data models of computation, leading toward the final project

- Understand the fundamentals of private data analysis

- Demonstrate awareness of common algorithmic techniques, through scribe notes

Secondary Goals

- Describe the motivation and statement of central data science problems, measured by the midterm presentation (practice reading and presenting technical papers)
- Work in various big data models of computation, leading toward the final project (practice thinking about research!)
- Understand the fundamentals of private data analysis
- Demonstrate awareness of common algorithmic techniques, through scribe notes (familiarity with LaTeX)

Grading

- LaTeX summary of lectures 20%

- Midterm presentation 35%

- Final project 45%

Related Coursework

- CSCE 689: Special Topics on Algorithms for Big Data
- Taught by Professor Crawford
- MWF 10:20-11:00 am, HRBB 126

- Topics:
  - Streaming algorithms
  - Parallel algorithms
  - Sublinear time algorithms
  - Sketching algorithms

Useful Background

- Big Oh notation, e.g., $O(\log^{10} n)$, $O(\sqrt{n})$, $O(n^2)$
- Reductions, e.g., NP-hardness
- Mathematical maturity, exposure to reading and writing proofs

Questions?

Probability Basics

- Random variable (X)

- Sample space (Ω): Set of possible values (discrete/continuous, finite/infinite)

- Probability: $\Pr[X = x]$ represents the probability that the random variable X achieves value $x \in \Omega$

Joint and Conditional Probability

- Joint distribution: $\Pr[X = x, Y = y]$ is the probability X and Y achieve values x and y respectively

- Conditional distribution: $\Pr[X = x | Y = y]$ is the probability that X achieves the value x when Y achieves the value y

$$\Pr[X = x | Y = y] = \frac{\Pr[X = x, Y = y]}{\Pr[Y = y]}$$

- Marginal distribution: $\Pr[X = x] = \sum_{y \in \Omega_Y} \Pr[X = x | Y = y]$

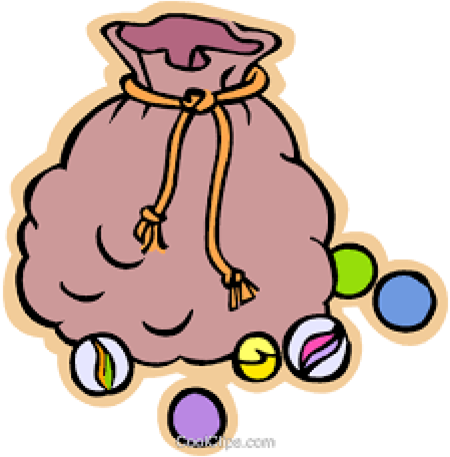
Independence

- Random variables X and Y are independent if $\Pr[X = x] = \Pr[X = x | Y = y]$ for all possible outcomes $x \in \Omega_X, y \in \Omega_Y$

Independence

- Suppose we have a bag with 1 red marble and 1 blue marble.
  - We draw a marble randomly from the bag
  - We put the marble back in the bag
  - We randomly draw another marble from the bag

- Let X be the color of the first marble drawn
- Let Y be the color of the second marble drawn



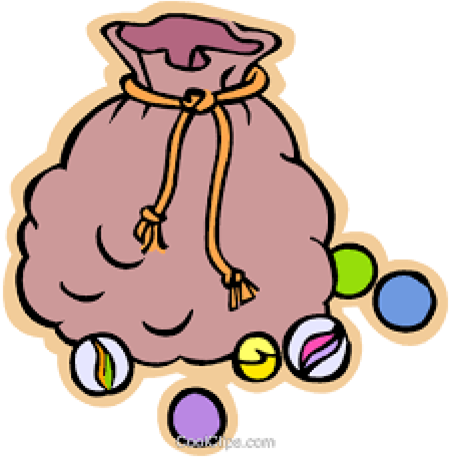
- Are X and Y independent?

Independence

- Suppose we have a bag with 1 red marble and 1 blue marble.
  - We draw a marble randomly from the bag
  - We DO NOT put the marble back in the bag
  - We randomly draw another marble from the bag

- Let X be the color of the first marble drawn
- Let Y be the color of the second marble drawn

- Are X and Y independent?



Boole's Inequality (Union Bound)

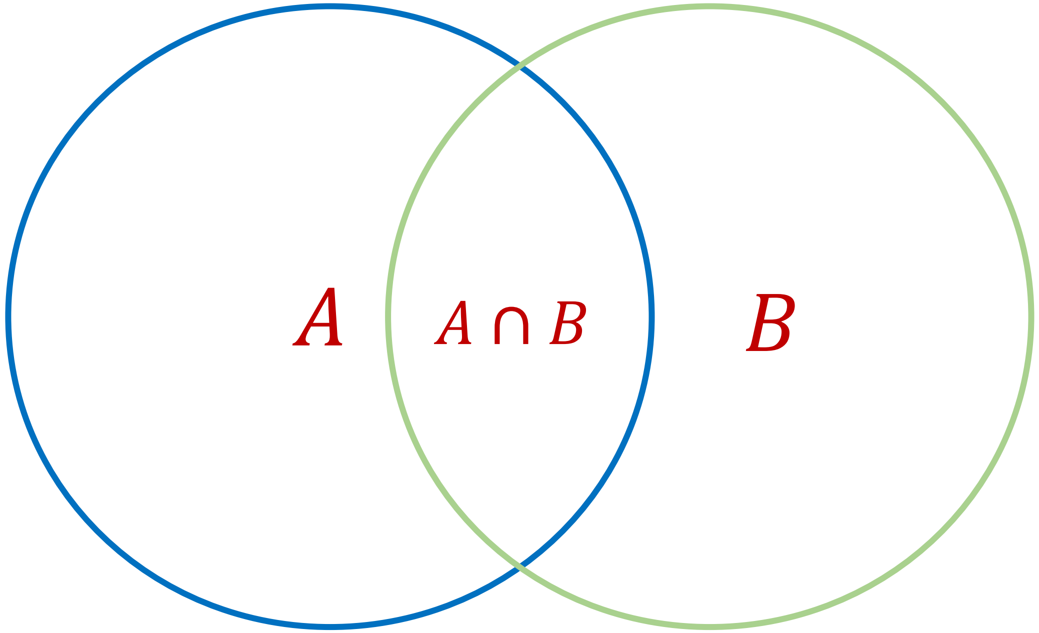
- Let S_1,…, S_k be a set of events that occur with probability p_1,…, p_k

- The probability that at least one of the events S_1,…, S_k occurs is at most $p_1 + \cdots + p_k$

- Implication: the probability that NONE of the events S_1,…, S_k occur is at least $1 - (p_1 + \cdots + p_k)$

Boole's Inequality (Union Bound)

- $\Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A \cap B]$



- Proof by induction

CSCE 689: Special Topics in Modern Algorithms for Data Science

Week 1: Probability basics

Samson Zhou

Trivia Question #1 (Birthday Paradox)

- Suppose we have a fair n-sided die. “On average”, how many times should we roll the die before we see a repeated outcome among the rolls? Example: 1, 5, 2, 4, 5

- $\Theta(1)$
- $\Theta(\log n)$
- $\Theta(\sqrt{n})$
- $\Theta(n)$

Trivia Question #2 (Limits)

- Let $c > 0$ be a constant. What is $\lim_{n \to \infty} \left(1 - \frac{c}{n}\right)^n$?

- 0
- $\frac{1}{c}$
- $\frac{1}{2c}$
- $\frac{1}{e^c}$
- 1

Trivia Question #3 (Coupon Collector)

- Suppose we have a fair n-sided die. “On average”, how many times should we roll the die before we all possible outcomes among the rolls? Example: 1, 5, 2, 4, 1, 3, 1, 6 for $n = 6$

- $\Theta(n)$
- $\Theta(n \log n)$
- $\Theta(n\sqrt{n})$
- $\Theta(n^2)$

Trivia Question #4 (Max Load)

- Suppose we have a fair n-sided die that we roll n times. “On average”, what is the largest number of times any outcome is rolled? Example: $1, 5, 2, 4, 1, 3, 1$ for $n = 7$

- $\Theta(1)$
- $\tilde{\Theta}(\log n)$
- $\tilde{\Theta}(\sqrt{n})$
- $\tilde{\Theta}(n)$

Birthday Paradox

- Suppose we have a room with 367 people. What is the probability that two people share the same birthday?

Birthday Paradox

- Suppose we have a room with 367 people. What is the probability that two people share the same birthday?

- Suppose we have a room with 23 people. What is the probability that two people share the same birthday?

Birthday Paradox

- Suppose we have a fair n-sided die that we roll $k = 1, 2, 3, 4$,... times. What is the probability we see a repeated outcome among the rolls?

Birthday Paradox

- Suppose we have a fair n-sided die that we roll $k = 1, 2, 3, 4$,... times. What is the probability we see a repeated outcome among the rolls?

$$\left(1 - \frac{0}{n}\right)$$

Birthday Paradox

- Suppose we have a fair n-sided die that we roll $k = 1, 2, 3, 4$,... times. What is the probability we see a repeated outcome among the rolls?

$$\left(1 - \frac{0}{n}\right)\left(1 - \frac{1}{n}\right)$$

Birthday Paradox

- Suppose we have a fair n-sided die that we roll $k = 1, 2, 3, 4$,... times. What is the probability we see a repeated outcome among the rolls?

$$\left(1 - \frac{0}{n}\right)\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)$$

Birthday Paradox

- Suppose we have a fair n-sided die that we roll $k = 1, 2, 3, 4$,... times. What is the probability we see a repeated outcome among the rolls?

$$\left(1-\frac{0}{n}\right)\left(1-\frac{1}{n}\right)\dots\left(1-\frac{k-1}{n}\right)$$

Birthday Paradox

- Suppose we have a fair n-sided die that we roll $k = 1, 2, 3, 4$,... times. What is the probability we see a repeated outcome among the rolls?

$$\left(1 - \frac{0}{n}\right)\left(1 - \frac{1}{n}\right)\ldots\left(1 - \frac{k-1}{n}\right) < \frac{1}{2} \qquad \text{for} \qquad k = O(\sqrt{n})$$

Birthday Paradox

- Suppose we have a fair n-sided die. “On average”, how many times should we roll the die before we see a repeated outcome among the rolls?
- $O(\sqrt{n})$

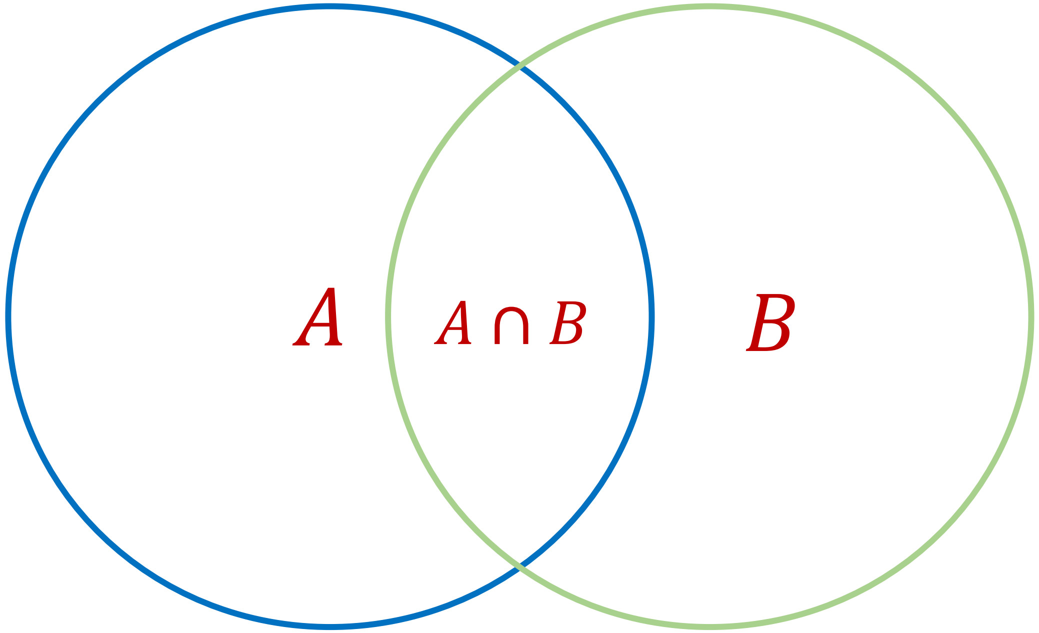
- But is it $\Theta(\sqrt{n})$?

Boole's Inequality (Union Bound)

- Let S_1,…, S_k be a set of events that occur with probability p_1,…, p_k
- The probability that at least one of the events S_1,…, S_k occurs is at most $p_1 + \cdots + p_k$
- Implication: the probability that NONE of the events S_1,…, S_k occur is at least $1 - (p_1 + \cdots + p_k)$

Boole’s Inequality (Union Bound)

- $\Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A \cap B]$



- Proof by induction

Birthday Paradox

- Suppose we have a fair n-sided die that we roll $k = 1, 2, 3, 4$,... times. What is the probability we see a repeated outcome among the rolls?

- Let S_i be the event that the i-th roll is a repeated outcome, conditioned on the previous rolls not being a repeated outcome
- $\Pr[S_i] = \frac{i-1}{n}$
- $\Pr[S_1 \cup \cdots \cup S_k] \leq \frac{0}{n} + \ \ldots + \frac{k-1}{n} \leq \frac{k^2}{n}$

Birthday Paradox

- Suppose we have a fair n-sided die. “On average”, how many times should we roll the die before we see a repeated outcome among the rolls?
- $\Theta(\sqrt{n})$

Hashing

- Suppose we have a data of images, how do we name them consistently?

Expected Value

- The expected value of a random variable X over Ω is:

$$\mathrm{E}[X] = \sum_{x \in \Omega} \Pr[X = x] \cdot x$$

- The “average value of the random variable"

- Linearity of expectation: $\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y]$

Expected Value

- Suppose we roll a 6-sided die

- Let X be the outcome of the roll

- What is $\mathrm{E}[X]$?

Moments

- For $p > 0$, the p-th moment of a random variable X over Ω is:

$$\mathrm{E}[X^p] = \sum_{x \in \Omega} \Pr[X = x] \cdot x^p$$

Variance

- The variance of a random variable X over Ω is:

$$\mathrm{Var}[X] = \mathrm{E}[X^2] - (\mathrm{E}[X])^2$$

- Linearity of variance for *independent* random variables: $\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$

- “How far numbers are from the average”

Variance

- Suppose X takes the value 1 with probability $\frac{1}{2}$ and takes the value -1 with probability $\frac{1}{2}$

- What is $\mathrm{E}[X]$?

- What is $\mathrm{Var}[X]$?

Variance

- Suppose Y takes the value 100 with probability $\frac{1}{2}$ and takes the value -100 with probability $\frac{1}{2}$

- What is $\mathrm{E}[Y]$?

- What is $\mathrm{Var}[Y]$?

Chebyshev's Inequality

- Let X be a random variable with expected value $\mu \coloneqq \mathrm{E}[X]$ and variance $\sigma^2 \coloneqq \mathrm{Var}[X]$

$$\Pr[|X-\mu| \geq k\sigma] \leq \frac{1}{k^2}$$

- "What is the probability a random variable is far away from its average?"