

Algorithms for Big-Data Management

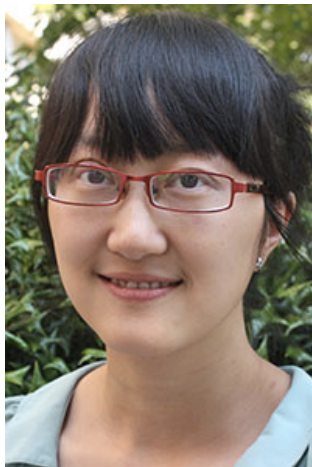
CompSci 590.04

Instructor: Ashwin Machanavajjhala

Course Staff



Ashwin Machanavajjhala
(Instructor)
Office Hours: By Appointment



Xi He
(Teaching Assistant)
Office Hours: Wed 4:30 – 6 PM

Administrivia

<http://www.cs.duke.edu/courses/fall15/compsci590.4/>

- Wed/Fri 3:05 – 4:20 PM
- “Reading Course + Project”
 - No exams!
 - Every class based on 1 (or 2) assigned papers that students *must* read.
- Projects: (50% of grade)
 - Individual or groups of size 3-4
- Assignments: (30% of grade)
 - There will be 3 assignments
- Class Participation: (other 20%)

Administrivia

- Projects: (50% of grade)
 - Ideas will be posted in the coming weeks
- Goals:
 - Literature review
 - Some original research/implementation
- Timeline (details will be posted on the website soon)
 - Sep 25: Choose Project (ideas will be posted ... new ideas welcome)
 - Oct 2: Project proposal (1-4 pages describing the project)
 - Oct 30: Mid-project review (2-3 page report on progress)
 - Nov 20: Final presentations and submission (6-10 page conference style paper + 15minute talk)

Why you should take this course?

- Industry, academic and government research identifies the value of analyzing large data collections in all walks of life.
 - *“What Next? A Half-Dozen Data Management Research Goals for Big Data and Cloud”, Surajit Chaudhuri, Microsoft Research*
 - *“Big data: The next frontier for innovation, competition, and productivity”, McKinsey Global Institute Report, 2011*

Why you should take this course?

- Very active field and tons of interesting research.
We will read papers in:
 - *Databases*
 - *Distributed Systems*
 - *Theory*
 - *Machine Learning*
 - *Privacy/Security*
 - ...

Why you should take this course?

- Intro to research by working on a cool project
 - *Read scientific papers*
 - *Formulate a problem*
 - *Perform a scientific evaluation*

Today

- Course overview
- An algorithm for sampling

INTRODUCTION

What is Big Data?

WHAT IS BIG DATA?

VOLUME

Large amounts of data.

VELOCITY

Needs to be analyzed **quickly**.

VARIETY

Different types of structured and unstructured data.

WHAT ARE THE VOLUMES OF DATA THAT WE ARE SEEING TODAY?



30 billion pieces of content were added to Facebook this past month by 600 million plus users.



Zynga processes **1 petabyte of content** for players every day; a volume of data that is unmatched in the social game industry.



More than 2 billion videos were watched on YouTube... yesterday.



The average teenager sends **4,762 text messages** per month.



32 billion searches were performed last month... on Twitter.

Source: Gartner

WHAT DOES THE FUTURE LOOK LIKE?

Worldwide IP traffic will **quadruple by 2015.**



By 2015, nearly **3 billion people**



will be online, pushing the data created and shared to nearly **8 zettabytes.**

<http://visual.ly/what-big-data>

Key questions enterprises are asking about Big Data:

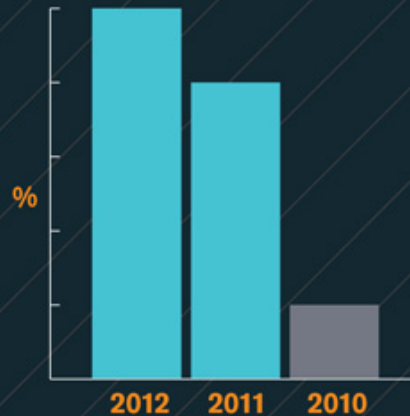
How to store and protect big data?

How to backup and restore big data?

How to organize and catalog the data that you have backed up?

How to keep costs low while ensuring that all the critical data is available when you need it?

Everyday business and consumer life creates **2.5 quintillion** bytes of data per day.

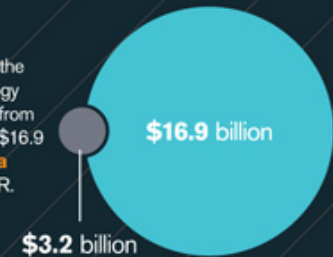


90% of the data in the world today has been created in the last two years alone.

Source: EIU

HOW IS THE MARKET FOR BIG DATA SOLUTIONS EVOLVING?

A new IDC study says the market for big technology and services will grow from \$3.2 billion in 2010 to \$16.9 billion in 2015. **That's a growth of 40% CAGR.**



58% of respondents expect their companies to increase spending on server backup solutions and other big data-related initiatives within the next three years.

Source: Economist Business Unit

A stylized map of North America in a light blue color, positioned behind the text.

2/3rds of surveyed businesses in North America said big data will become a concern for them within the next five years.

Source: Economist Business Unit

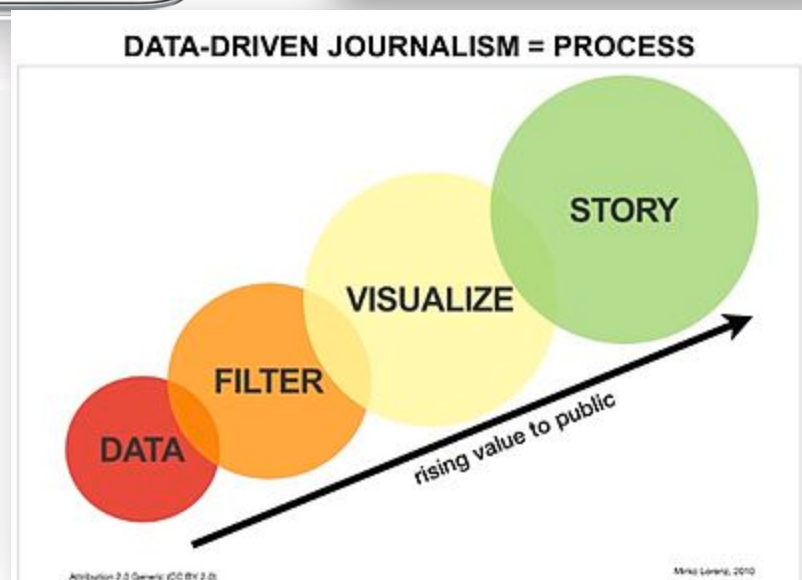
<http://visual.ly/what-big-data>

3 Key Trends

- Increased data collection
- (Shared nothing) Parallel processing frameworks on commodity hardware
- Powerful analysis techniques at both the population and individual levels

Big-Data impacts all aspects of our life

The collage illustrates the integration of big data into daily life. It features a Yahoo! search page with various news and search results, a Facebook page for 'Celebs on Facebook' showing a public figure's profile and posts, and a smartphone displaying the 'INSTANT ECG' app by iANESTHESIA LLC, which shows a heart rate monitor interface.



The value in Big-Data ...

My Yahoo! | Make Y! your homepage | Feb 23, 2009 | Hi, Michael | Enjoying the sunshine | Sign Out | Page Options

Web | Images | Video | Local | Shopping | More

YAHOO! Web Search

MY FAVORITES + Add

- Yahoo Sites
- Mall (3)
- Weather (72°)
- Finance (Dow)
- Sports (2)
- Movies (2)
- Horoscope
- eBay (2)
- Local
- USAToday
- NY Times
- Shopping
- Facebook (12)
- OMG
- Y! Buzz
- Messenger (9)

RECOMMENDED

- Netflix
- Wired
- Amazon

Sights to see before you die

A globetrotting travel writer names 10 unforgettable spots that belong on your list. Places of immense beauty

- Tips for finding cheap airfare
- Great family destinations
- Find vacation package deals

TOP SEARCHES « Prev | Next »

- Blagojevich
- vtv
- The Biggest Loser
- Oprah Winfrey
- Oil Prices
- Jay Leno
- Jesse Jackson Jr
- Robert Pattinson
- Casey Anthony
- Twilight

More Top Searches

THE ALL-NEW 2010 RX
REINVENTING THE VEHICLE THAT INVENTED IT ALL.

click to explore Visit lexus.com - Ad Feedback

SPOTLIGHT « Prev | Next »

Most Popular Soup Recipes FOOD

- Big Batch Vegetable Soup
- Chicken Noodle Soup
- Butternut Squash Soup
- Broccoli Chowder
- Pumpkin Soup

More Soup Recipes

Markets: Dow: 8,514 +20.34% Nasdaq: 1,248.62 -10.34%

enter ticker symbol Get Quotes SPONSORED BY: Scottrade

Recommended links

+79% clicks
vs. randomly selected

Personalized
News Interests

+250% clicks
vs. editorial one size fits all

Top Searches

+43% clicks
vs. editor selected

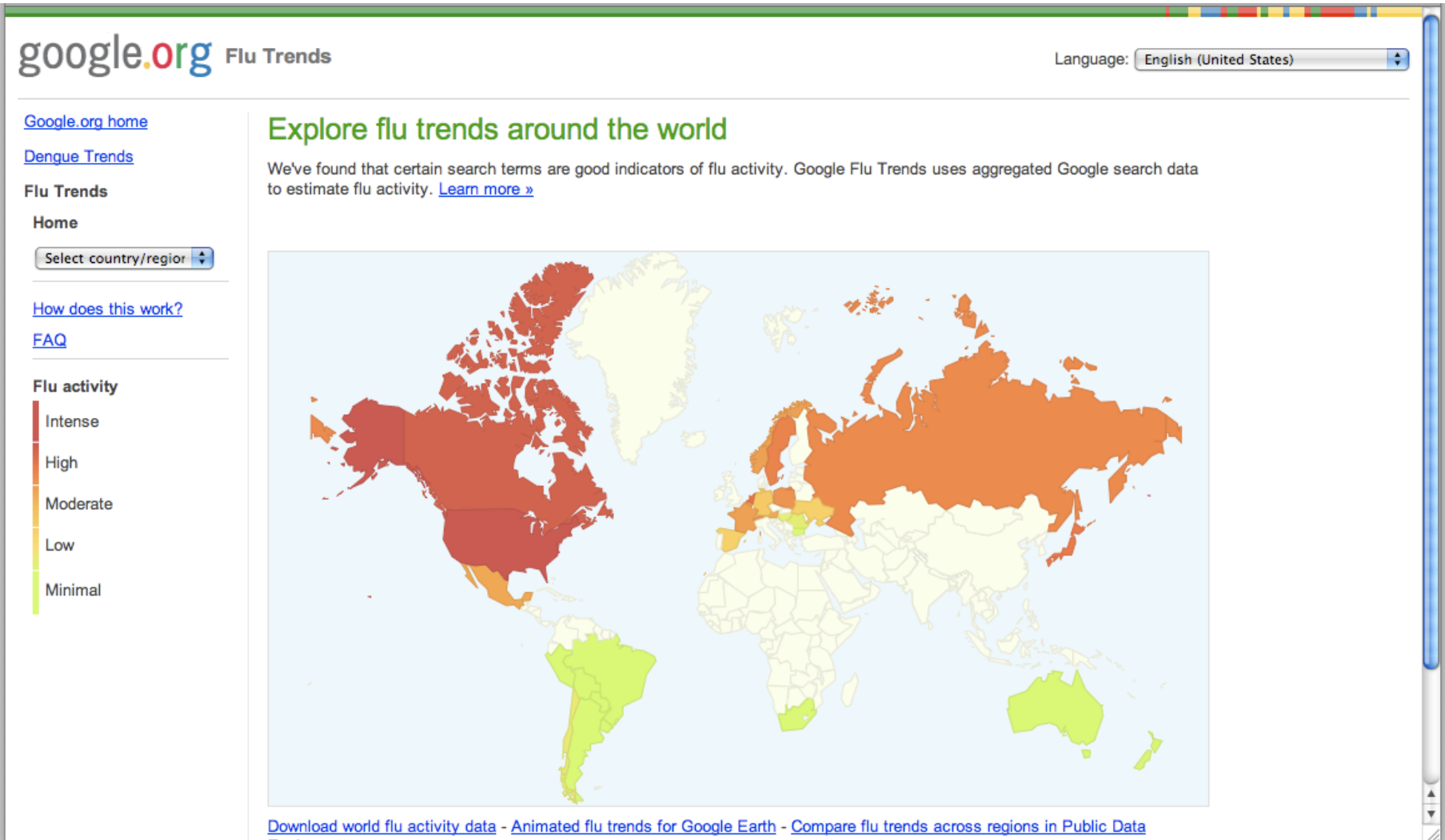
The value in Big-Data ...

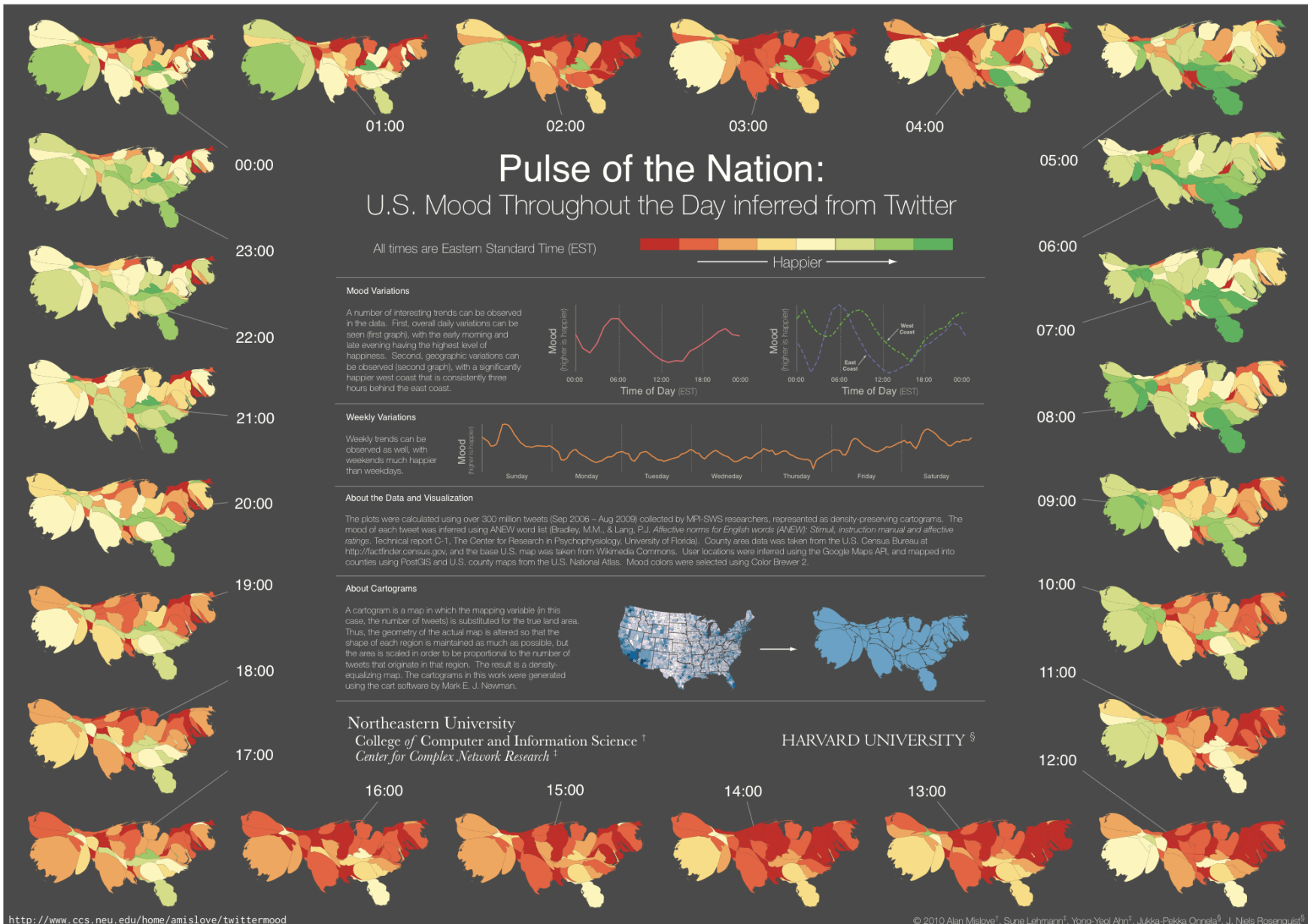


“
*If US healthcare were to use **big data** creatively and effectively to drive efficiency and quality, the sector could create more than*
”
\$300 billion in value every year.

McKinsey Global Institute Report

Example: Google Flu





<http://www.ccs.neu.edu/home/amislove/twittermood/>

Course Overview

We will learn strategies for handling data that is ...

1. large
2. fast
3. sensitive
4. partitioned

... and along the way we will learn a number of useful tricks.

Course Overview

Strategy 1: Compute approximate answers on *large data*

- Sampling
 - Reservoir Sampling
 - Sampling with indices/Joins
 - Monte Carlo method
 - Markov Chains

Course Overview

Strategy 2: Compute approximate answers *on fast data*

- Streaming
 - Sketches
 - Online Aggregation
 - Online learning

Course Overview

Strategy 3: Throw a lot of hardware at *large data*

- Parallel Architectures & Algorithms
 - Map Reduce
 - Graph processing architectures : Bulk Synchronous parallel and asynchronous models
 - (Graph connectivity, Matrix Multiplication, Belief Propagation)

Course Overview

Strategy 4: Add noise to handle *sensitive data*

- Computing under noise
 - Differential privacy
 - Histograms
 - Range queries
 - Sorting

Course Overview

Strategy 5: Join *partitioned data*

- Joining datasets & Record Linkage
 - Theta Joins: or how to optimally join two large datasets
 - Clustering similar documents using minHash
 - Correlation Clustering

SAMPLING

Why Sampling?

- Approximately compute quantities when
 - Processing the entire dataset takes too long.
How many tweets mention Obama?
 - Computation is intractable
Number of satisfying assignments for a DNF.
 - Do not have access or expensive to get access to entire data.
How many restaurants does Google know about?
Number of users in Facebook whose birthday is today.
What fraction of the population has the flu?

Zero-One Estimator Theorem

Input: A universe of items U (e.g., all tweets)
A subset G (e.g., tweets mentioning Obama)

Goal: Estimate $\mu = |G|/|U|$

Algorithm:

- Pick N samples from U $\{x_1, x_2, \dots, x_N\}$
- For each sample, let $Y_i = 1$ if $x_i \in G$.
- Output: $Y = \sum Y_i/N$

Theorem: Let $\epsilon < 1.5$. If $N > (1/\mu) (3 \ln(2/\delta)/\epsilon^2)$, then
 $\Pr[(1-\epsilon)\mu < Y < (1+\epsilon)\mu] > 1-\delta$

Zero-One Estimator Theorem

Algorithm:

- Pick N samples from $U \{x_1, x_2, \dots, x_N\}$
- For each sample, let $Y_i = 1$ if $x_i \in G$.
- Output: $Y = \sum Y_i / N$

Theorem: Let $\varepsilon < 1.5$. If $N > (1/\mu) (3 \ln(2/\delta)/\varepsilon^2)$, then

$$\Pr[(1-\varepsilon)\mu < Y < (1+\varepsilon)\mu] > 1-\delta$$

Proof: Homework

Estimating multiple properties

- Suppose there are 'm' subsets of interest G_1, G_2, \dots, G_m
- Goal: Estimate $\mu_i = |G_i|/|U|$ for all i
- How many samples do we need?

Estimating multiple properties

- Suppose there are 'm' subsets of interest G_1, G_2, \dots, G_m
- Goal: Estimate $\mu_i = |G_i|/|U|$ for all i
- How many samples do we need?
- Answer: $N > (3/\mu\epsilon^2) (\ln m + \ln(2/\delta))$, where $\mu = \min_i \mu_i$

Simple Random Sample

- Given a table of size N , pick a subset of n rows, such that each subset of n rows is equally likely.
- How to sample n rows?
- ... if we don't know N ?

Reservoir Sampling

Highlights:

- Make one pass over the data
- Maintain a reservoir of n records.
- After reading t rows, the reservoir is a simple random sample of the first t rows.

Reservoir Sampling [Vitter ACM ToMS '85]

Algorithm R:

- Initialize reservoir to the first n rows.
- For the $(t+1)^{\text{st}}$ row R ,
 - Pick a random number m between 1 and $t+1$
 - If $m \leq n$, then replace the m^{th} row in the reservoir with R

Proof

Proof

- If $N = n$, then $P[\text{row is in sample}] = 1$. Hence, reservoir contains all the rows in the table.
- Suppose for $N = t$, the reservoir is a simple random sample. That is, each row has n/t chance of appearing in the sample.
- For $N = t+1$:
 - $(t+1)$ st row is included in the sample with probability $n/(t+1)$
 - Any other row:
$$P[\text{row is in reservoir}] = P[\text{row is in reservoir after } t \text{ steps}] * P[\text{row is not replaced}]$$
$$= n/t * (1 - 1/(t+1)) = n/(t+1)$$

Complexity

- Running time: $O(N)$
- Number of calls to random number generator: $O(N)$
- Expected number of elements that may appear in the reservoir:
$$n + \sum_{t=0}^{N-1} n/(t+1) = n(1 + H_N - H_n) \approx n(1 + \ln(N/n))$$
- Is there a way to sample faster? in time $O(n(1 + \ln(N/n)))$??

Faster algorithm

- Algorithm R skips over (does not insert into reservoir) a number of records ($N - n(1 + \ln(N/n))$)
- At any step t , let $S(n,t)$ denote the number of rows skipped by the Algorithm R.
 - Involved $O(S)$ time and $O(S)$ calls to the random number generator.
- $P[S(n,t) = s] = ?$

Faster algorithm

- At any step t , let $S(n,t)$ denote the number of rows skipped by the Algorithm R.
- $P[S(n,t) = s] =$ for all $t < x \leq t+s$, row x was not inserted into reservoir, but row $t+s+1$ is inserted.
 $= \{ 1 - n/(t+1) \} \times \{ 1 - n/(t+2) \} \times \dots \times \{ 1 - n/(t+s) \} \times n/(t+s+1)$
- We can derive expression for CDF:
 $P[S(n,t) \leq s] = 1 - (t/t+s+1)(t-1/t+s)(t-2/t+s-1) \dots (t-n+1/t+s-n+2)$

Faster Algorithm

Algorithm X

- Initialize reservoir with first n rows.
- After seeing t rows, randomly sample a skip $s = S(n,t)$ from the CDF
- Pick a number m between 1 and n
- Replace the m th row in the reservoir with the $(t+s+1)$ st row.
- Set $t = t + s + 1$

Faster Algorithm

Algorithm X

- Initialize reservoir with first n rows.
- After seeing t rows, randomly sample a skip $s = S(n,t)$ from the CDF
 - Pick a random U between 0 and 1
 - Find the minimum s such that $P[S(n,t) \leq s] \leq 1-U$
- Pick a number m between 1 and n
- Replace the m th row in the reservoir with the $(t+s+1)$ st row.
- Set $t = t + s + 1$

Algorithm X

- Running time:
Each skip takes $O(s)$ time to compute
Total time = sum of all the skips = $O(N)$
- Expected number of calls to the random number generator
= $2 * \text{expected number of rows in the reservoir}$
= $O(n(1 + \ln(N/n)))$ **optimal!**

See paper for algorithm which has optimal runtime

Summary

- Sampling is an important technique for computation when data is too large, or the computation is intractable, or if access to data is limited.
- Reservoir sampling techniques allow computing a sample even without knowledge of the size of the data.
 - Also can do weighted sampling [Efraimidis, Spirakis IPL 2006]
- Very useful for sampling from streams (e.g., twitter stream)

References

- J. Vitter, “Random Sampling with a Reservoir”, ACM Transaction on Mathematical Software, 1985
- P. Efraimidis, P. Spirakis, “Weighted random sampling with a reservoir”, Journal Information Processing Letters, 97(5), 2006
- R. Karp, R. Luby, N. Madras, “Monte Carlo Approximation Algorithms for Enumeration Problems”, Journal of Algorithms, 1989