Lecture 22: Differential Privacy

COSC 480 Data Science, Spring 2017

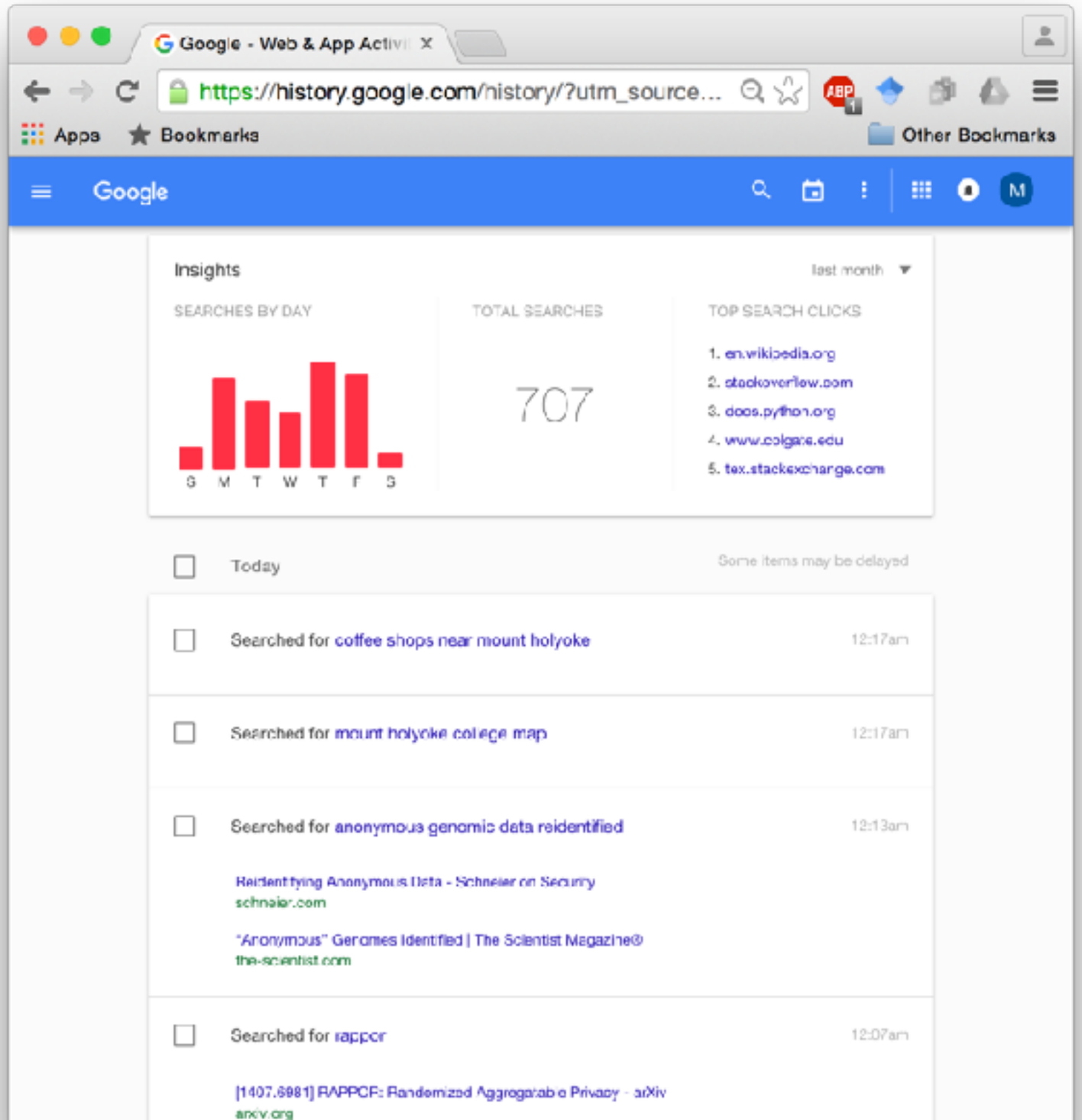
Michael Hay

Logistics

- Tonight: weekly status on project is due
- Tuesday (tomorrow):
  - lab time devoted to projects, visualization and/or lab 7
  - attendance encouraged but not required
- Wednesday: mini-presentations
- Friday: talk that was required is now optional

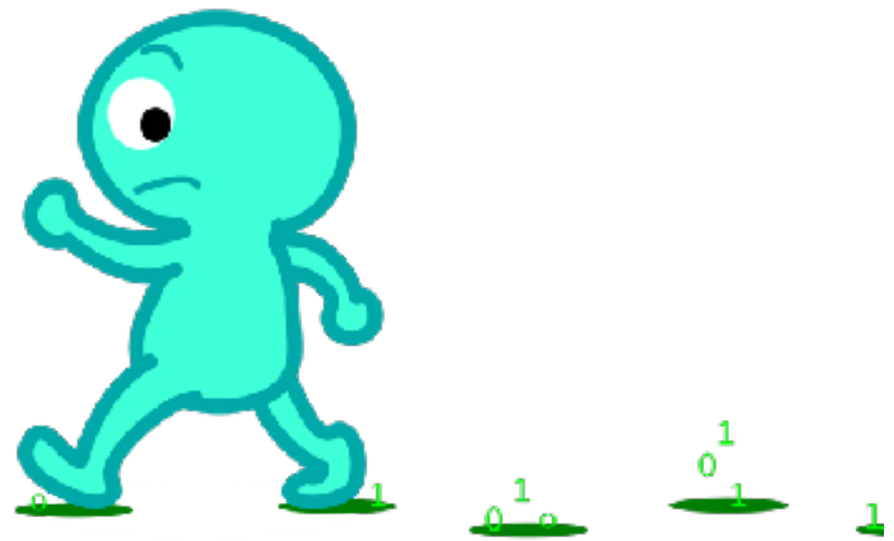
Thought experiment

- Summer internship at Google
- Gain access to search history of total stranger.
- What could you learn?

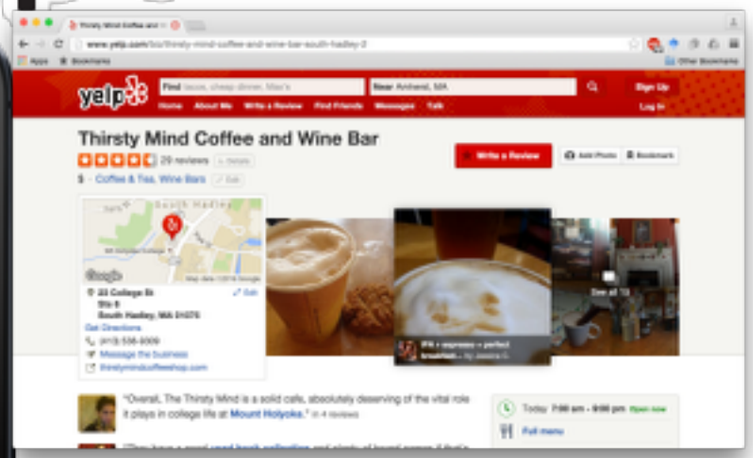


Big Data: volume and variety

- When we interact with digital technologies, we leave behind a digital trace



http://teachingprivacy.icsi.berkeley.edu/



Collected data can be sensitive

The Opinion Pages

Campaign Stops

How Racist Are We? Ask Google

By SETH STEPHENS-DAVIDOWITZ

Barack Obama won 52.9 percent of the popular vote in 2008 and 365 electoral votes, 95 more than he needed. Many naturally concluded that prejudice was not a major factor against a black presidential candidate in modern America. My research, a comparison of Americans' Google searches and their voting patterns, found otherwise. If my results are correct, racial animus cost Mr. Obama many more votes than we may have realized.

Doug Mills/The New York Times

Quantifying the effects of racial prejudice on voting is notoriously problematic. Few people admit bias in surveys. So I used a new tool, Google Insights, which tells researchers how often words are searched in different parts of the United States.

The New York Times

SundayReview | The Opinion Pages

WORLD U.S. N.Y. / REGION BUSINESS TECHNOLOGY SCIENCE HEALTH SPORTS OPINION

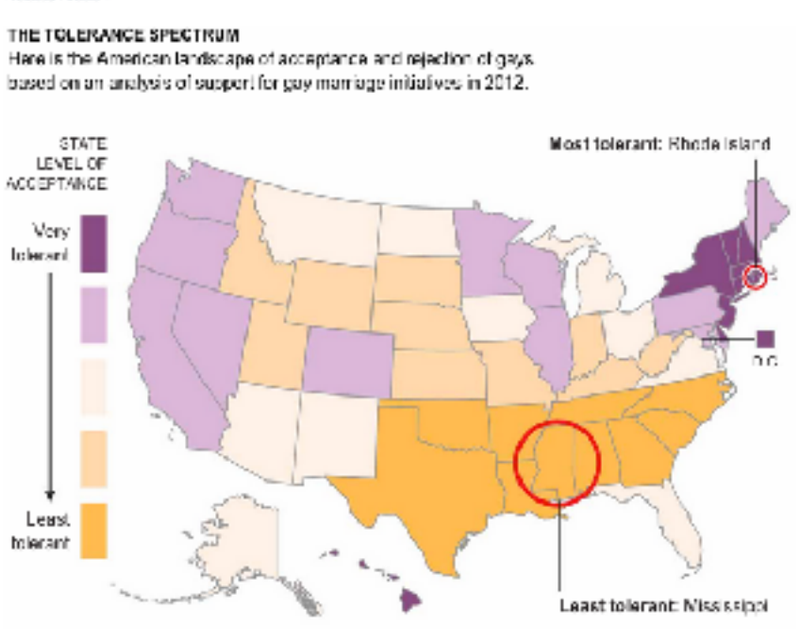
OPINION

Where the Closet Is Still Common

Related Article »

THE TOLERANCE SPECTRUM

Here is the American landscape of acceptance and rejection of gays, based on an analysis of support for gay marriage initiatives in 2012.



ARE THERE FEWER GAY MEN IN THE LEAST TOLERANT STATES?

The online public sphere gives that impression. Percent of male Facebook

But in the private sphere, the gay share is larger and

The New York Times

SundayReview | The Opinion Pages

WORLD U.S. N.Y. / REGION BUSINESS TECHNOLOGY SCIENCE HEALTH SPORTS OPINION

OPINION

Dr. Google Will See You Now

By SETH STEPHENS-DAVIDOWITZ

Published August 9, 2013

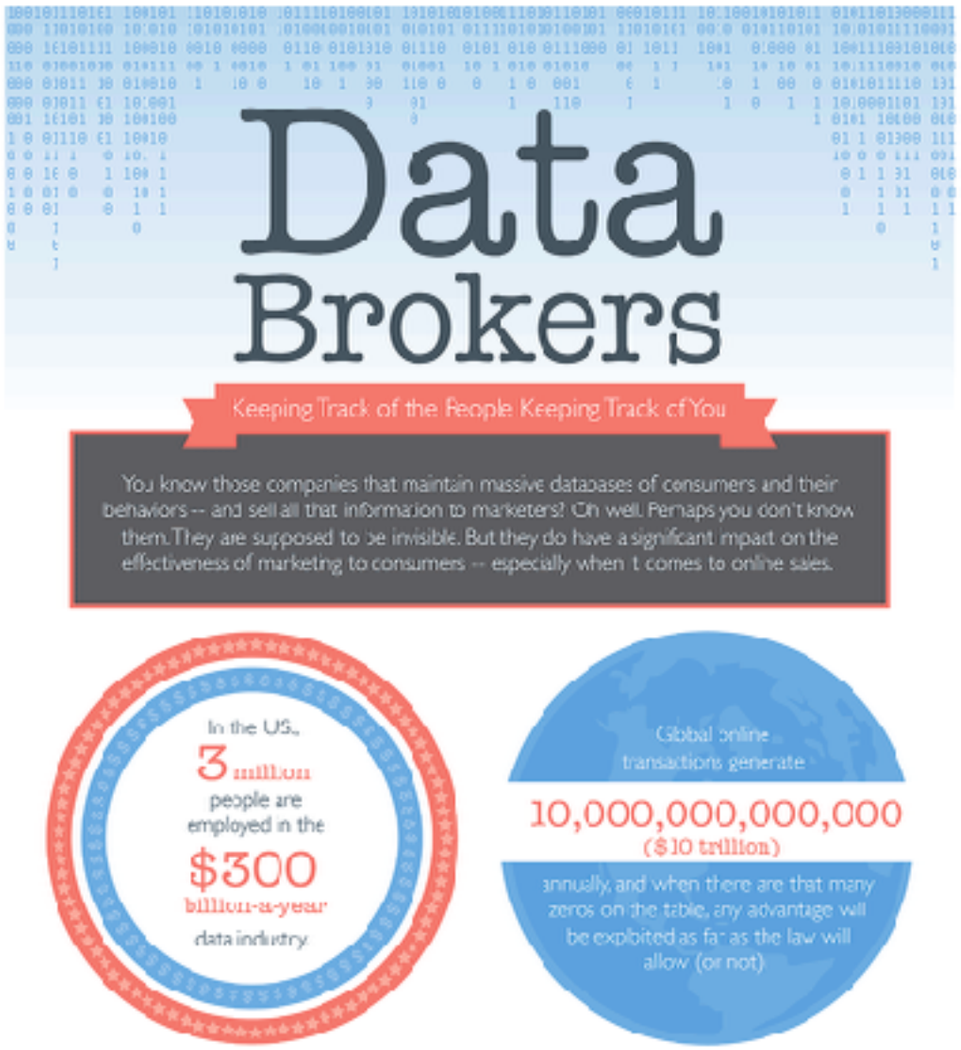
FEELING good today? You are not alone. We're approaching Aug. 11, which is, excluding Christmastime, the happiest day of the year in this country. Or at least the day with the lowest rate of depression, according to our Google searches.

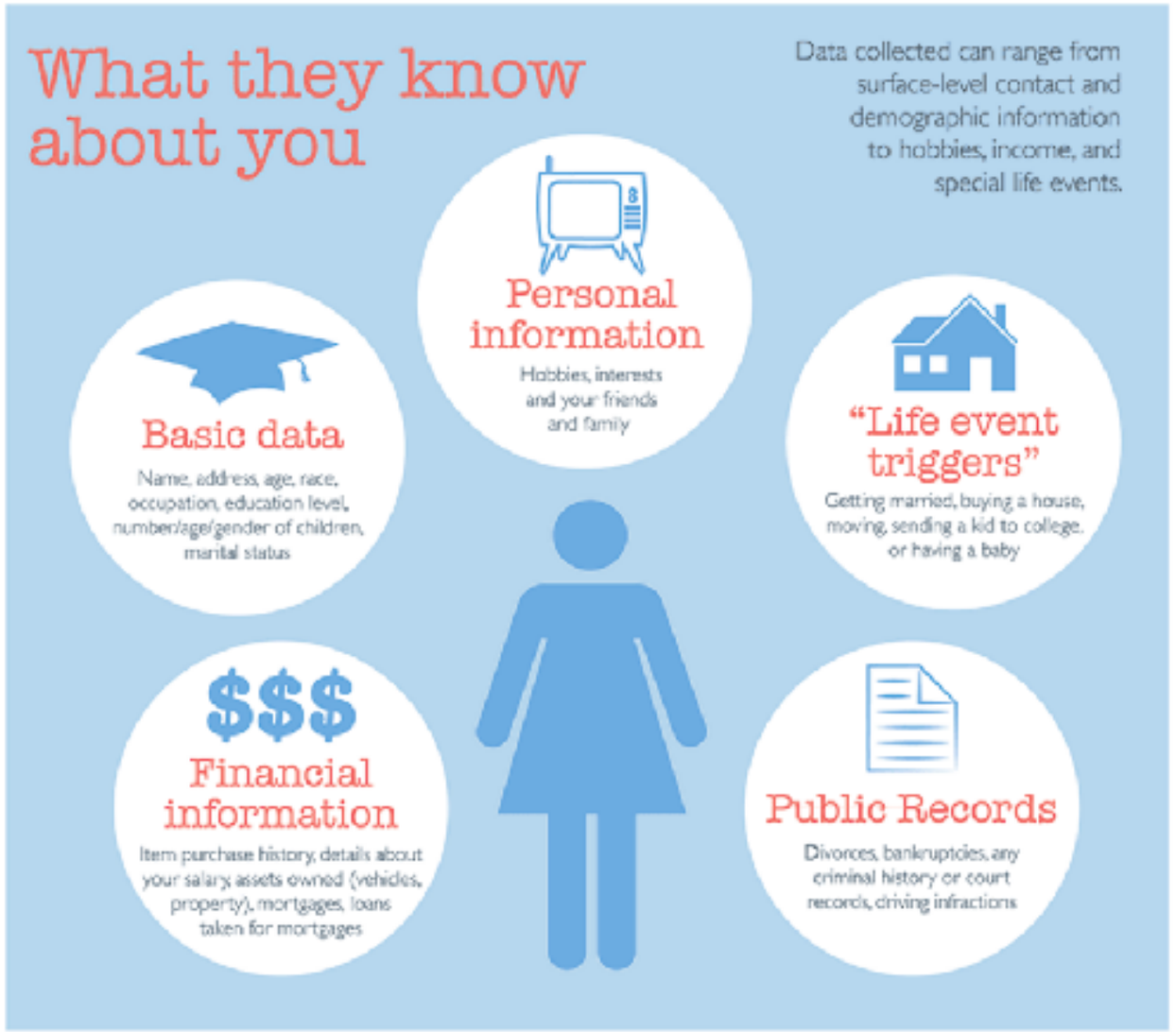
Edwige Tak Image

I recently explored what Google searches tell us about depression, by which I mean, loosely, dips in mood. Using anonymous, aggregate data from tens of millions of queries, I measured what proportion of searches included the word "depression" and were categorized by Google's algorithms as health related on each day of the year, across the United States, over a period of nine years. (This method counts searches like "depression symptoms" and "depression treatment" but not searches like "The Great Depression.") Not every health-related search using "depression" is a sign that someone is depressed, and not everyone who is depressed queries Google. But thanks to the incredibly large sample size, meaningful patterns emerge.

FACEBOOK
TWITTER
GOOGLE+
SAVE
EMAIL
SHARE
PRINT
REPRINTS

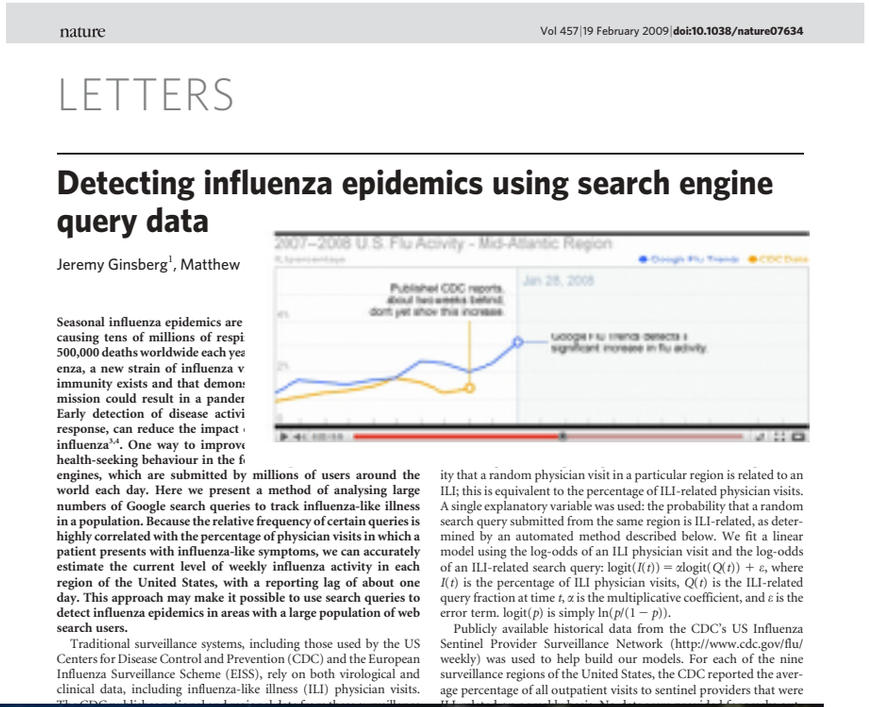
Decisions driven by data





http://visual.ly/data-brokers

Data science used for social good


nature Vol 457|19 February 2009|doi:10.1038/nature07634

LETTERS

Detecting influenza epidemics using search engine query data

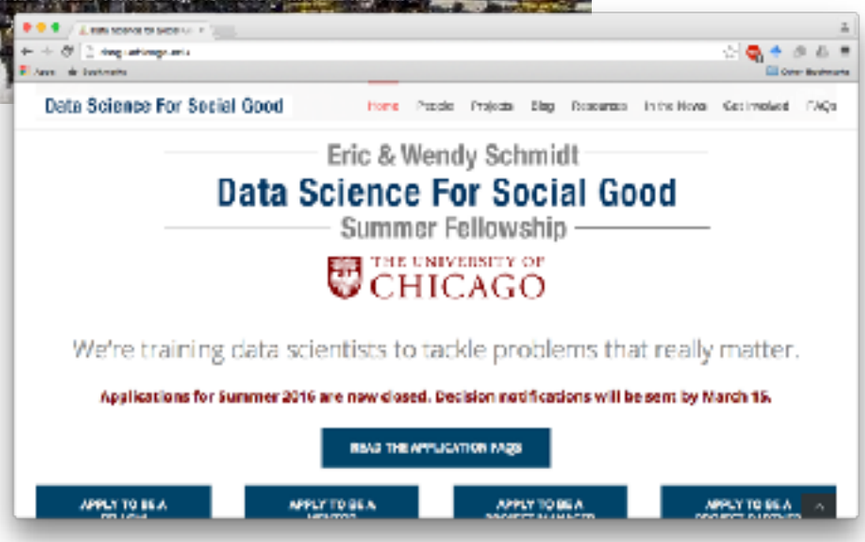
Jeremy Ginsberg[1], Matthew

Seasonal influenza epidemics are causing tens of millions of respi 500,000 deaths worldwide each yea enza, a new strain of influenza v immunity exists and that demons mission could result in a pander Early detection of disease activi response, can reduce the impact influenza[3,4]. One way to improve health-seeking behaviour in the fo engines, which are submitted by millions of users around the world each day. Here we present a method of analysing large numbers of Google search queries to track influenza-like illness in a population. Because the relative frequency of certain queries is highly correlated with the percentage of physician visits in which a patient presents with influenza-like symptoms, we can accurately estimate the current level of weekly influenza activity in each region of the United States, with a reporting lag of about one day. This approach may make it possible to use search queries to detect influenza epidemics in areas with a large population of web search users.

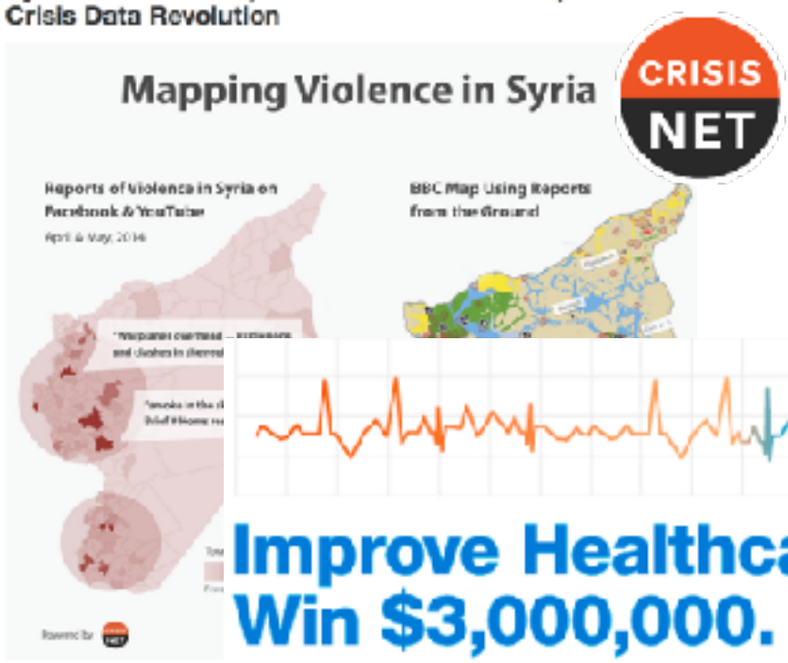
Traditional surveillance systems, including those used by the US Centers for Disease Control and Prevention (CDC) and the European Influenza Surveillance Scheme (EISS), rely on both virological and clinical data, including influenza-like illness (ILI) physician visits.

ity that a random physician visit in a particular region is related to an ILI; this is equivalent to the percentage of ILI-related physician visits. A single explanatory variable was used: the probability that a random search query submitted from the same region is ILI-related, as determined by an automated method described below. We fit a linear model using the log-odds of an ILI physician visit and the log-odds of an ILI-related search query: $\text{logit}(I(t)) = \alpha\text{logit}(Q(t)) + \varepsilon$, where $I(t)$ is the percentage of ILI physician visits, $Q(t)$ is the ILI-related query fraction at time t, α is the multiplicative coefficient, and ε is the error term. $\text{logit}(p)$ is simply $\ln(p/(1-p))$.

Publicly available historical data from the CDC's US Influenza Sentinel Provider Surveillance Network (http://www.cdc.gov/flu/weekly) was used to help build our models. For each of the nine surveillance regions of the United States, the CDC reported the average percentage of all outpatient visits to sentinel providers that were



Syrian Social Media, Journalists' Secret Weapon in the Crisis Data Revolution



Improve Healthcare, Win $3,000,000.

Identify patients who will be admitted to a hospital within the next year, using historical claims data.

Description Evaluation

Bits

More than 71 million individuals in the U American Hospital Association. Studies admissions. Is there a better way? Can Heritage Provider Network (HPN) believ

Researchers Draw Romantic Insights From Maps of Facebook Networks

By STEVE LOHR OCTOBER 26, 2013 8:00 AM 47 Comments

A graphical representation of one person's network neighborhood on Facebook. Cameron Marlow/Facebook

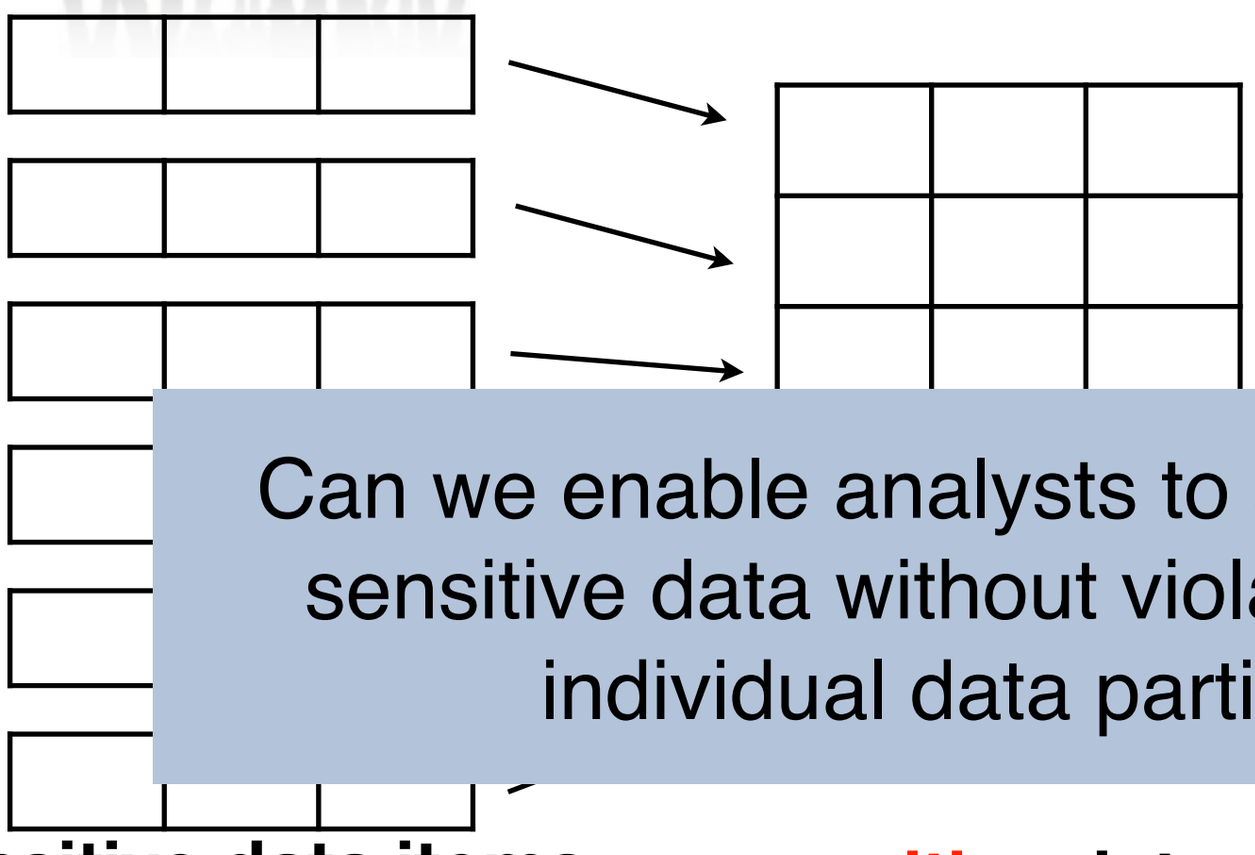
Problem setting

INDIVIDUALS

sensitive data items (e.g. health records)

COLLECTOR

(trusted)



sensitive data set

ANALYST

(untrusted)

✔

global properties

"The incidence of flu has doubled in last week"

"What factors make patient most susceptible to flu?"

facts

e flu"

Can we enable analysts to accurately study sensitive data without violating privacy of individual data participants?

Pitfalls of data anonymization

Published data (Massachusetts Group Insurance Commission)

name	ssn		gender	dob	zip	diagnosis
joe	123-56-763	4	male	1/4/64	01045	cancer
mary	113-36-425	2	female	3/24/45	01312	flu
						allergy

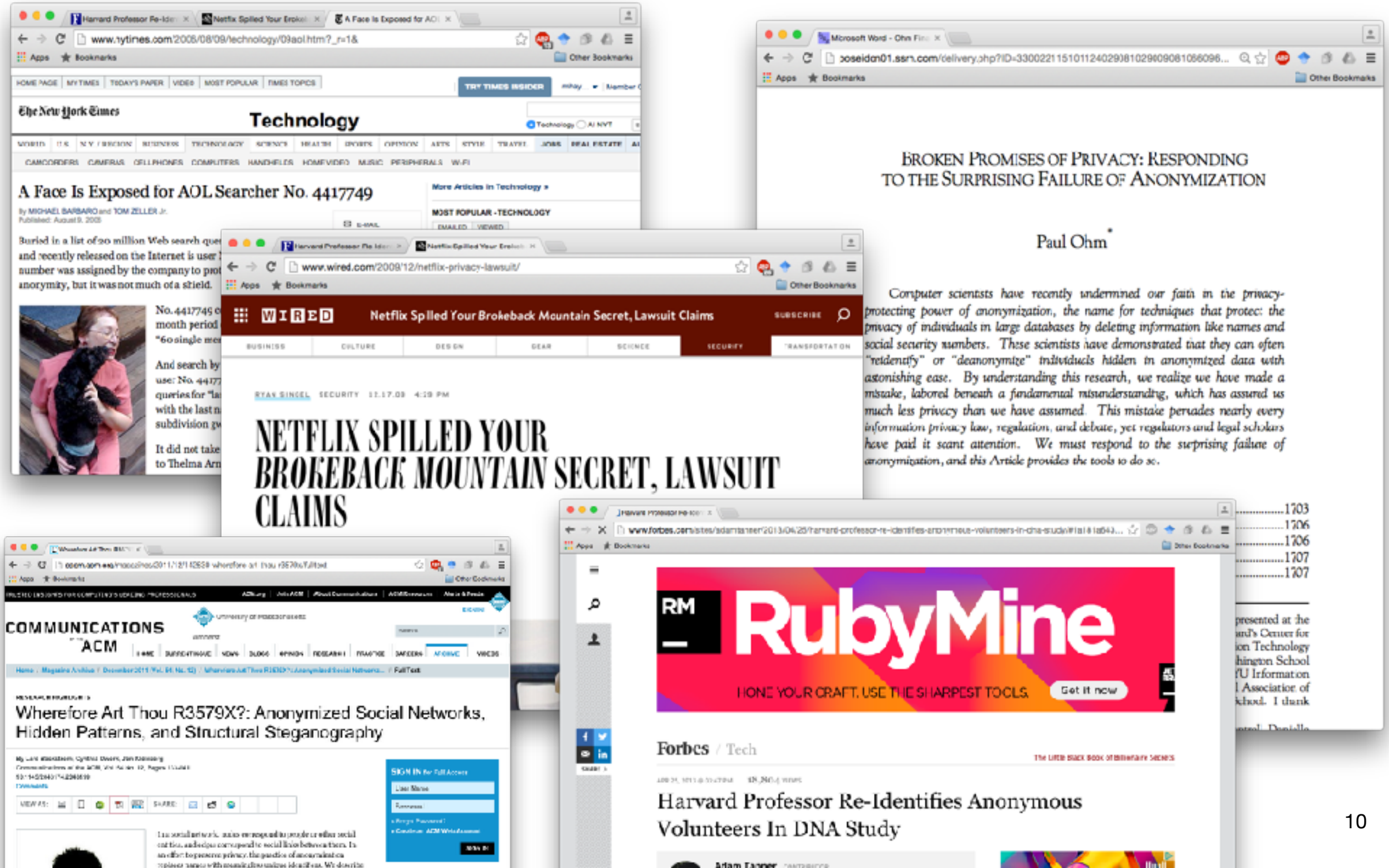
Sweeney re-identified the medical records of the governor of Massachusetts!

External

name	party	gender	dob	zip
bill	Rep	male	7/31/45	02138
mary	Rep	female	3/24/45	01312
joe	Dem	male	1/4/64	01045

[Sweeney 2002]

Anonymization failures



Outline

- Introduction
- Randomized response
- Differential privacy
- Preliminary work on trajectory data

[Warner, JASA 65]

Randomized response

- Pollster wants to conduct survey
- Survey topic is sensitive
    - Marijuana smoking habits
    - Political support for Trump
    - etc.
- How to elicit unbiased answers?

[Warner, JASA 65]

Randomized Response

1. Interviewer poses a sensitive question to respondent, but respondent does not answer aloud.

   Q1: Have you ever been to Flour & Salt?

2. Respondent flips *biased* coin (Let *p be* probability of heads).

   Outcome not revealed to interviewer.
   Associate Heads with Yes, and Tails with No

3. Respondent answers Question 2:

   Q2: Does the flipped coin agree with your answer to Q1?

Disclosure depends on bias of coin

Q2			
Y			
Y	N	T	(1-p)

- When $p = 1/2$, what does Q2 reveal about Q1? Nothing!
- When $p = 1$, what does Q2 reveal about Q1? Everything!
- What about when $p = 0$? What about $p \in (1/2, 1)$?

What can we learn using randomized response?

- Let θ denote (unknown) proportion of population who smoke pot.
- Sample n people from population. For each, do randomize response.
- Let X be the number of people in sample who answer yes to Q2.
- Given X, how to estimate θ?

Estimating θ

- Let P_{YES} denote the probability of saying yes to Q2

$$P_{YES} = P(\text{respondent smokes pot and flips H}) +$$

- Solve for θ

$$\theta = \frac{P_{YES} - (1-p)}{(2p-1)}$$

- Replace P_{YES} with X/n, the proportion who said yes in sample

$$\hat{\theta}(X) = \frac{X/n - (1-p)}{(2p-1)}$$

method of moments estimator

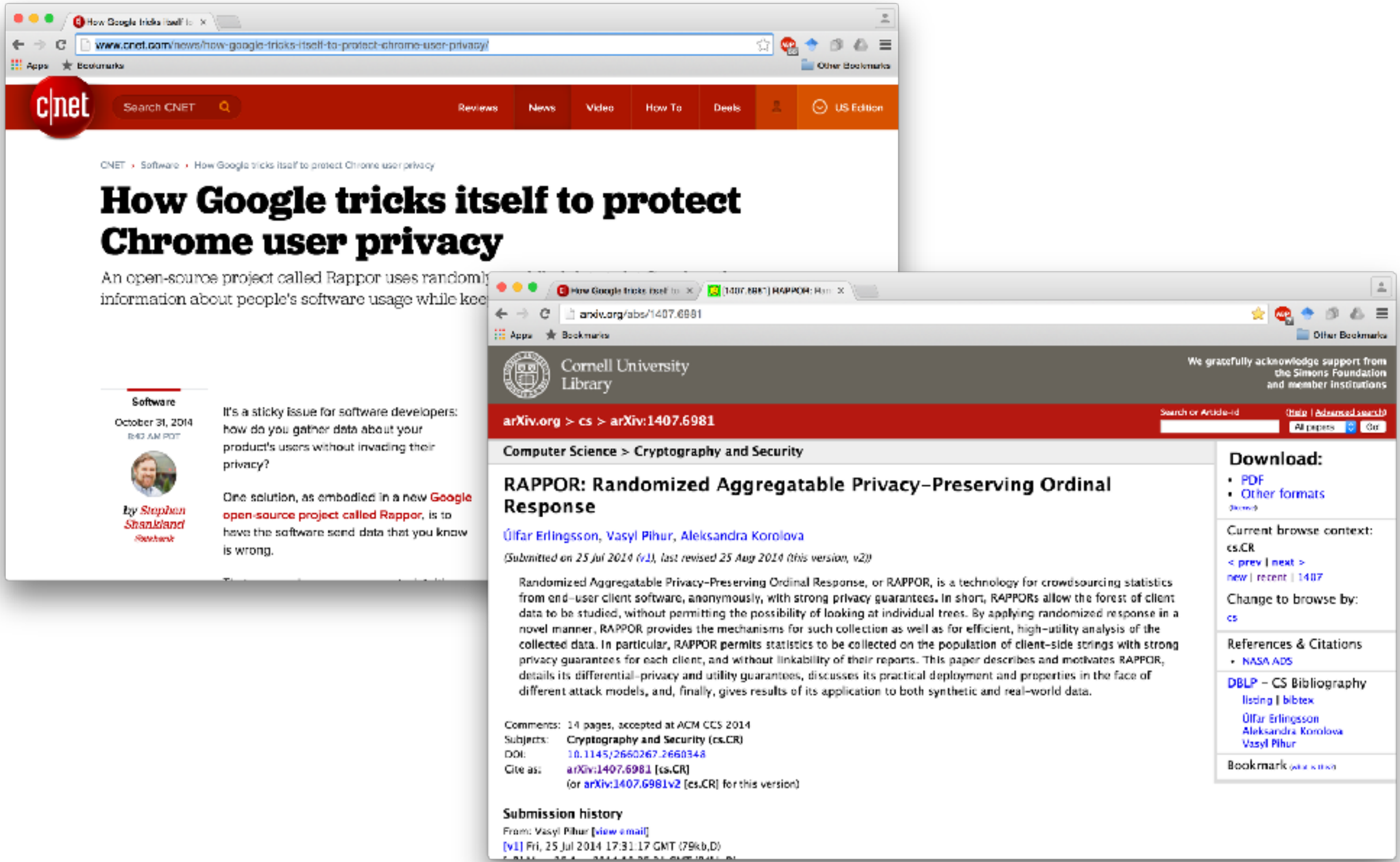
Error of estimator

- Expected squared error is error due to **sampling** plus error due to **randomized response**

$$\mathbb{E}\left[\left(\hat{\theta}(X)-\theta\right)^2\right] = \frac{\theta(1-\theta)}{n} + \frac{\frac{1}{16(p-1/2)^2}-\frac{1}{4}}{n}$$

- For p ≠ 1/2, as sample size *n* grows, error goes to zero.

Randomized response in action

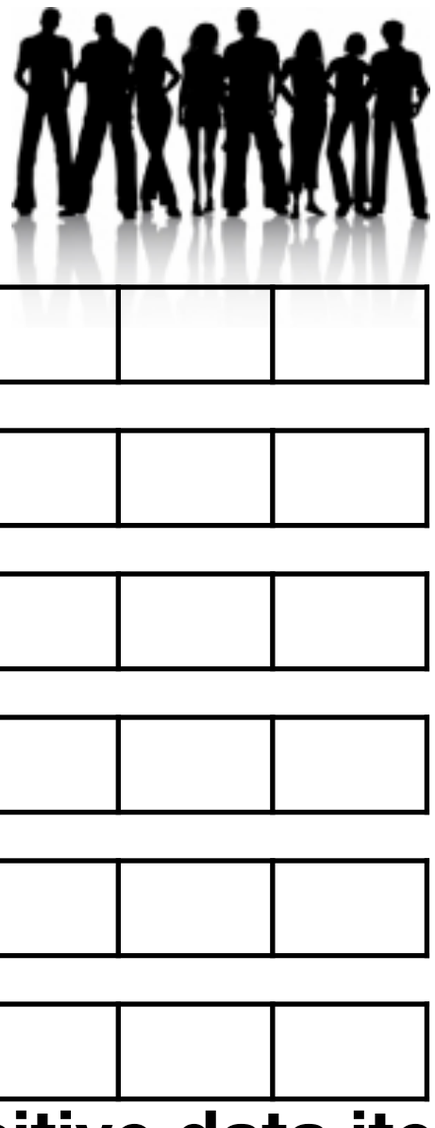


Outline

- Introduction
- Randomized response
- Differential privacy
- Preliminary work on trajectory data

Problem setting

INDIVIDUALS | **COLLECTOR** | **ANALYST**



(trusted)

(untrusted)

✔ **global properties**

"The incidence of flu has doubled in last week"

"What factors make patient most susceptible to flu?"

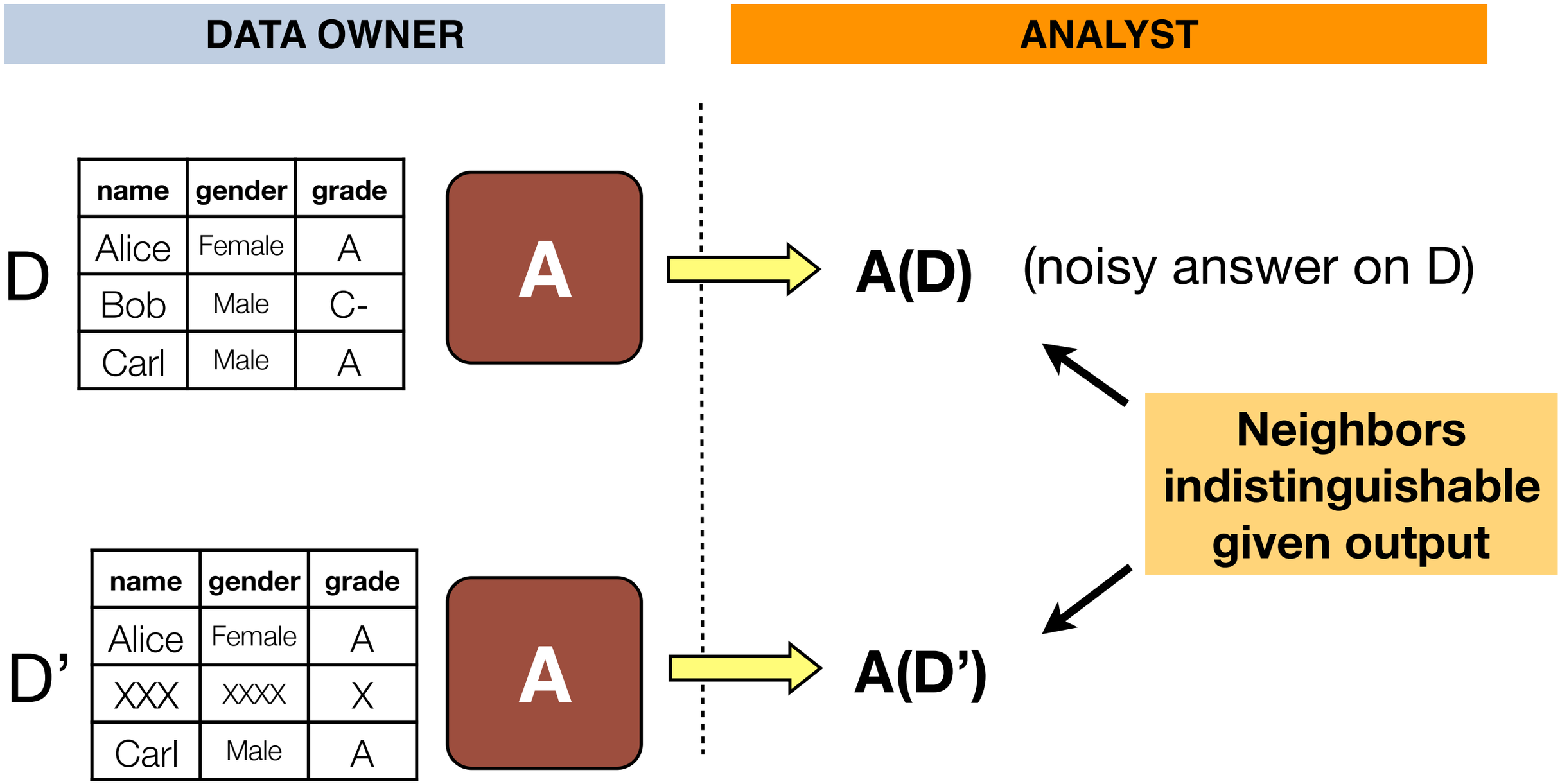
✖ **sensitive facts**

"Alice has the flu"

sensitive data items (e.g. health records)

sensitive data set

The differential guarantee



Two databases are **neighbors** if they differ by at most one tuple

[Dwork, TCC 06]

Differential privacy

A randomized algorithm A provides **ε-differential privacy** if:
for all neighboring databases D and D', and
for any set of outputs *S*:

$$Pr[\mathcal{A}(D) \in S] \leq e^{\epsilon} Pr[\mathcal{A}(D') \in S]$$

parameter, smaller means more privacy

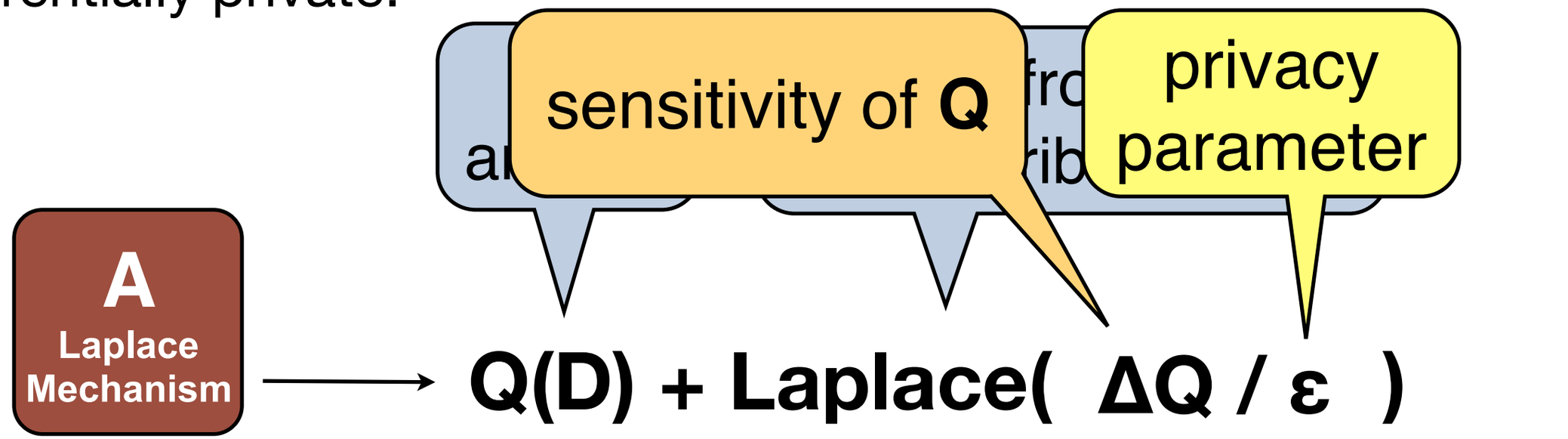
Informal: "Algorithm A is differentially private if the probability of a particular output is basically the same whether an individual's record is included or not."

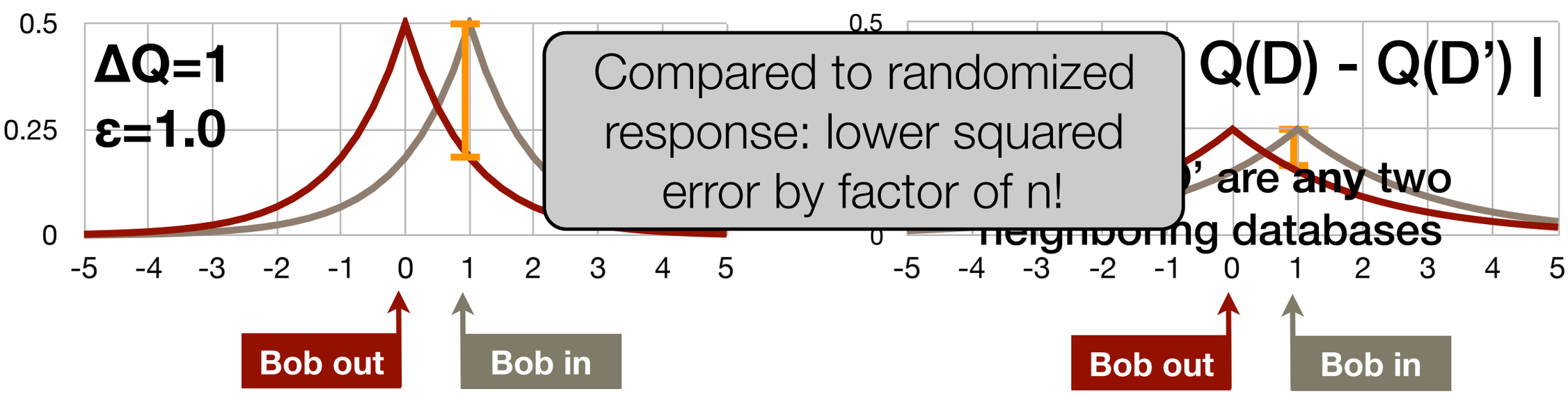
Randomized response satisfies ln(p/(1-p))-differential privacy

[Dwork, TCC 06]

The Laplace mechanism

The following algorithm for answering any statistical query **Q** is ε-differentially private:





Properties of privacy guarantee

- Suppose A_1 is ε_1-differentially private and A_2 is ε_2-differentially private

- ***Post-processing:*** additional computation on $A_1(D)$ does not affect privacy guarantee (not true for anonymization!)

- ***Sequential composition***: running both is $\varepsilon_1+\varepsilon_2$-differentially private

- ***Parallel composition***: partition D into *non-overlapping subsets* D_1 and D_2, $A_1(D_1)$ and $A_2(D_2)$ is $\max(\varepsilon_1,\varepsilon_2)$-differentially private

- Implication: complex algorithms from simpler building blocks

Research in differential privacy

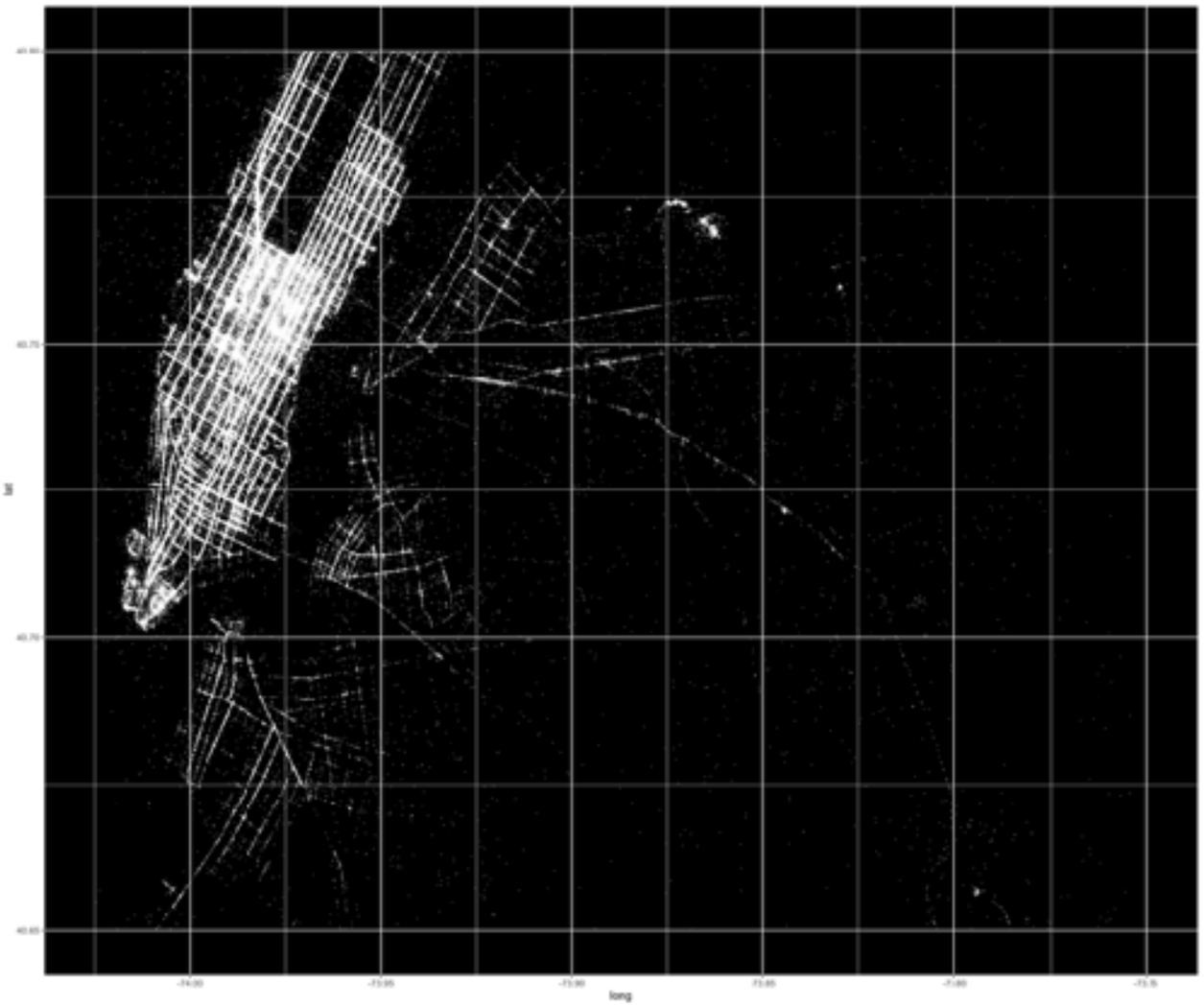
- Fundamental questions: how many queries can be accurately answered?
- Counting queries on tabular data have been intensively studied [PODS 10, PVLDB 10 , SIGMOD 11, PVLDB 14]
- Applications
  - Machine learning: regression, classification
  - Data mining: frequent itemset mining, sequence mining
  - Network data [ICDM 09, PODS 09]
  - Trajectory data
- Alternative (mostly weaker) privacy definitions [PODS 09, CRYPTO 12]

Outline

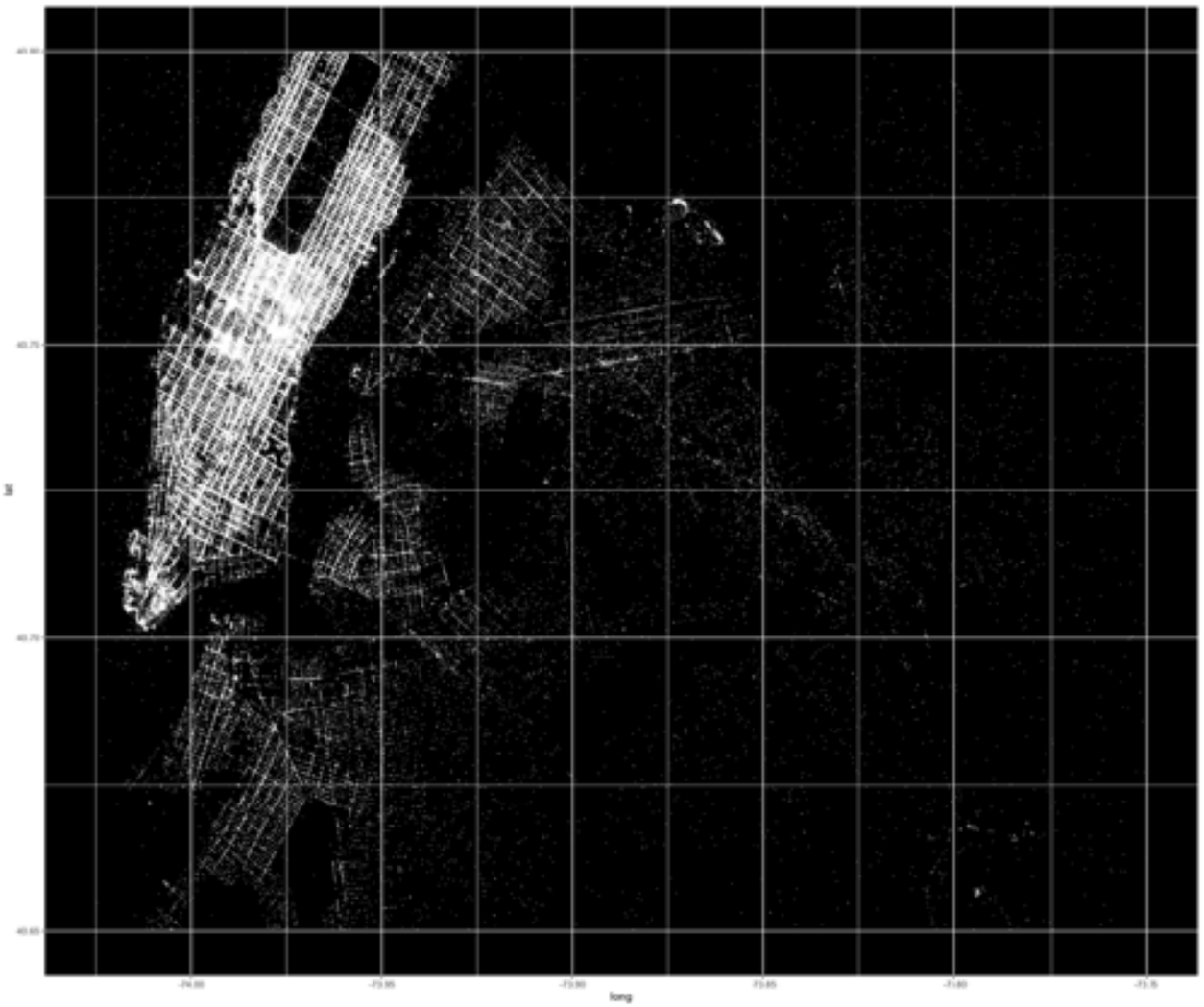
- Introduction
- Randomized response
- Differential privacy
- Preliminary work on trajectory data

Trajectory Data

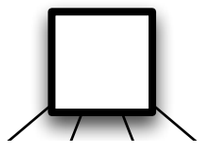
NYC Taxi Pick Ups
(2010, sample of 300K)

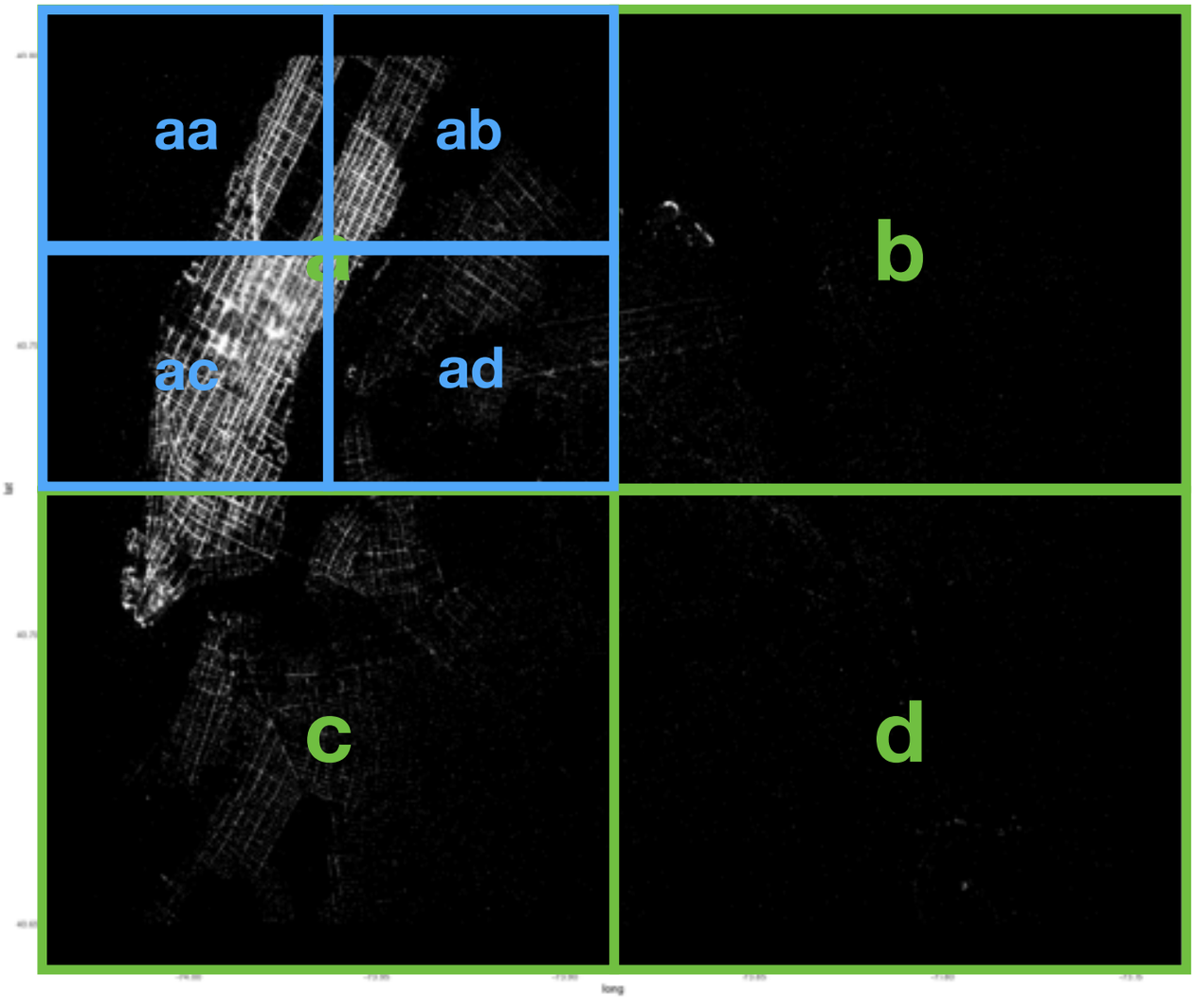


NYC Taxi Drop Offs
(2010, sample of 300K)



Quadtree

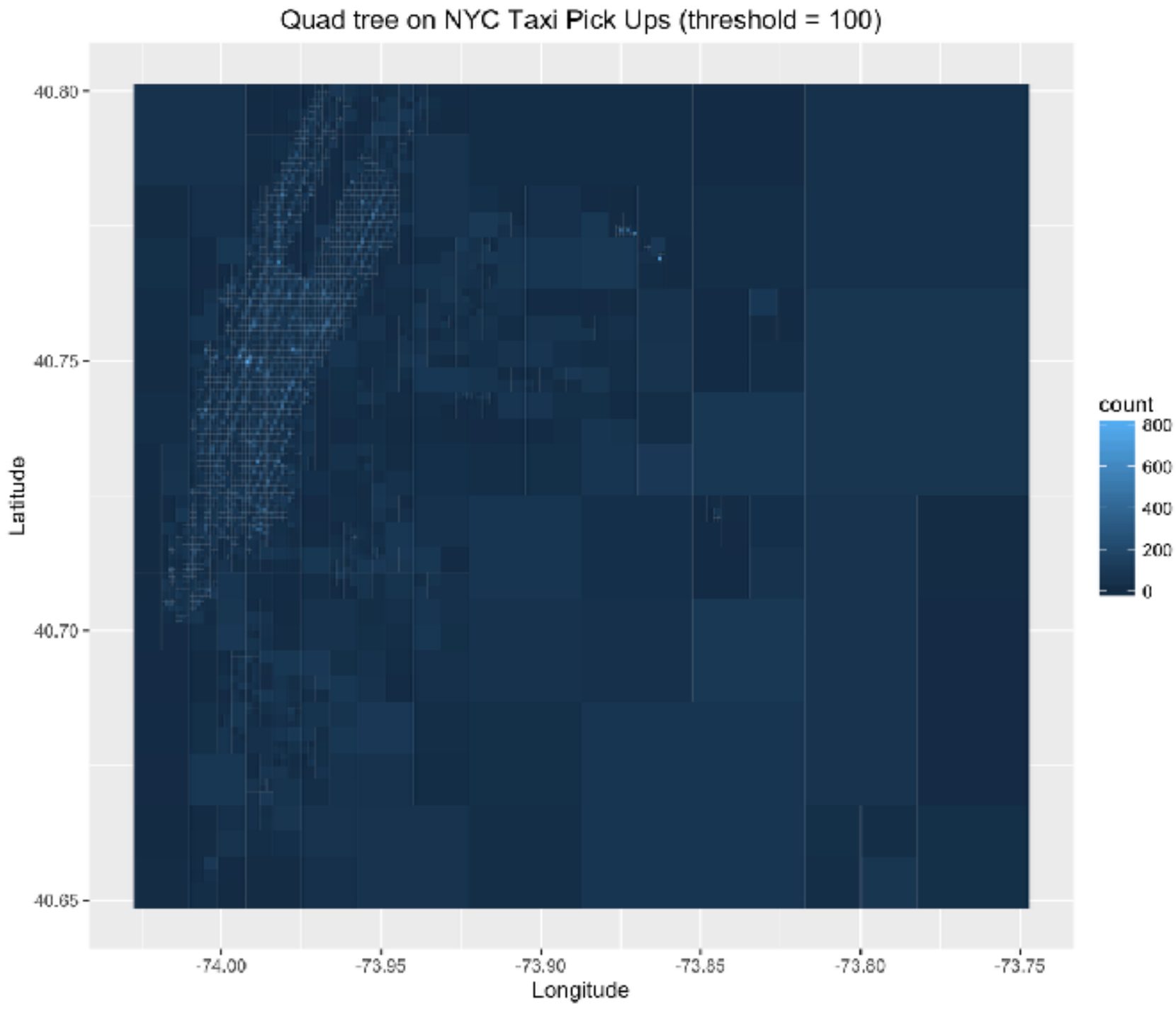




- Leaves of quad tree define a partition of 2d space
- Each leaf reports count of points in its region

Strategy: “drill down” on dense regions

- For node of quadtree, associate count of points in its region
- Split node as long as count is above some threshold θ



Baseline: Simple Tree

- let *T* be ***full*** quad tree to some height *h*
- for each node *v* in *T* (top to bottom)
    - compute noisy count for *v* (using Laplace mechanism)
    - if noisy count is below θ, ***prune*** sub-tree below *v*
- privacy
    - sensitivity of count is 1
    - siblings cover disjoint regions (parallel composition)
    - … **but** parent-child regions overlap (sequential composition)
- **problem**: scale of noise must be h/ε (grows with height *h*)