Medical Data Privacy and Ethics in the Age of Artificial Intelligence

Lecture 12: Privacy Risk Assessment and Facial Image Privacy Risks

Zhiyu Wan, PhD (wanzhy@shanghaitech.edu.cn)
Assistant Professor of Biomedical Engineering
ShanghaiTech University

April 16, 2025

Learning Objectives of This Lecture

- Understand the definition of Marketer risk
- Know where to find data (availability)
  - How to be a data detective
- Know the concept of facial image privacy

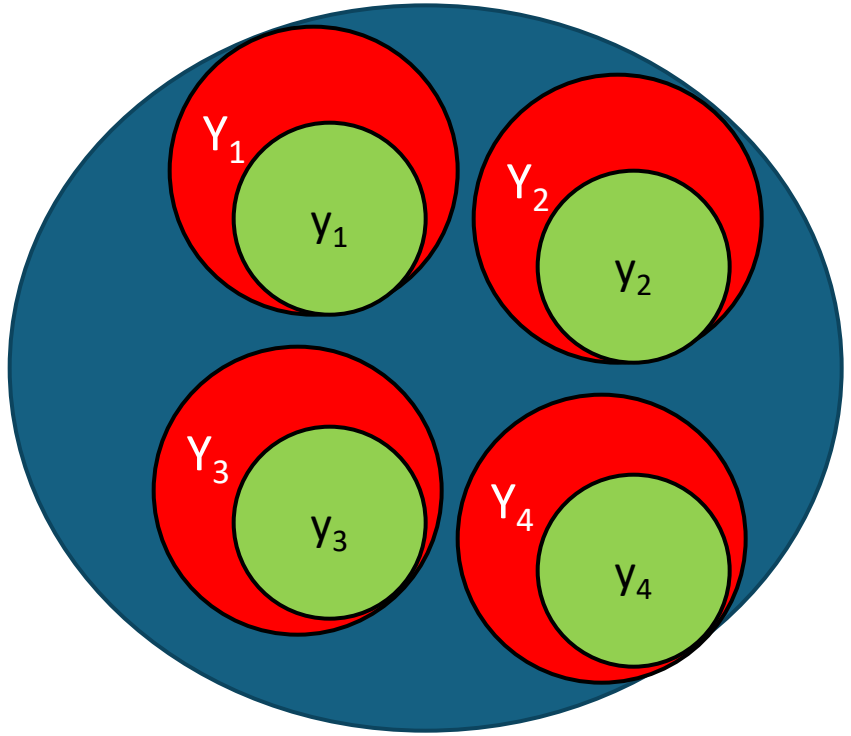
Risk Models

- Risk means something different to everyone
- Can be modeled in various ways*
  - Prosecutor← $1/\min_i(y_i)$
  - Journalist ← $1/\min_i(Y_i)$
  - Marketer ← $n^{-1}\sum_i \frac{y_i}{Y_i}$

y_i = size of group in sample

Y_i = size of group in population

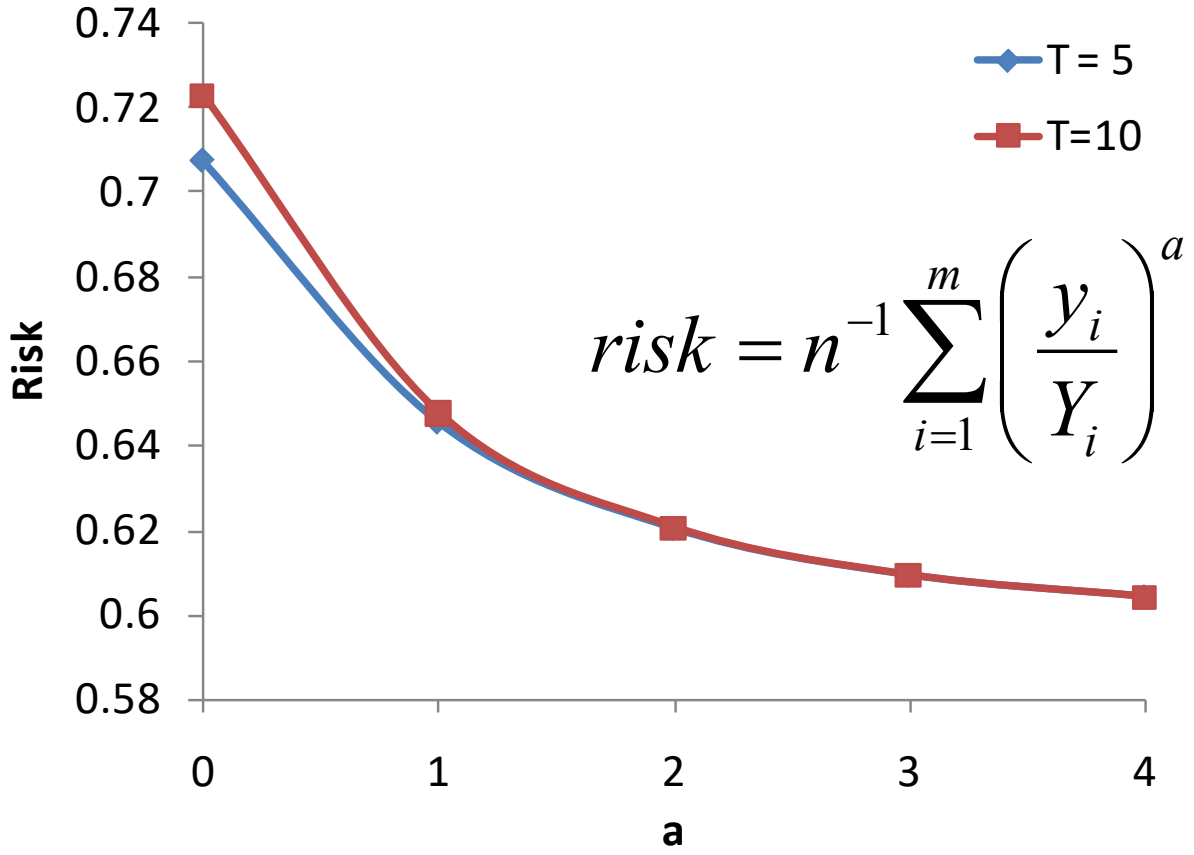
n = sample size



*F. Dankar and K. El Emam. A method for evaluating marketer re-identification risk. EDBT/ICDT Workshops 2010.

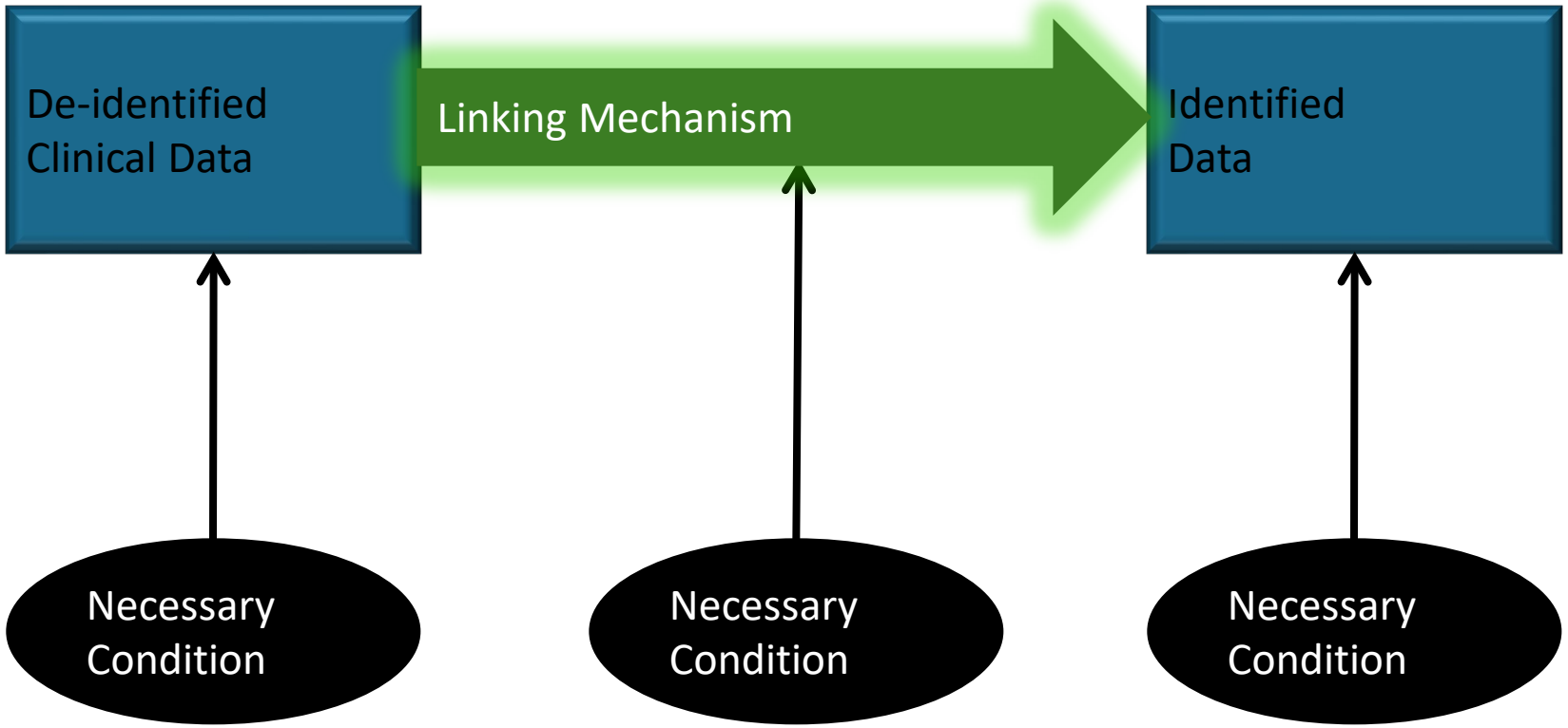
Tuning the Marketer Model

- State of Kentucky risk $\{DOB, Gender, County, Race\}$



K. Benitez and B. Malin. Evaluating re-identification risks with respect to the HIPAA Privacy Rule. JAMIA. 2010; 17: 169-177.

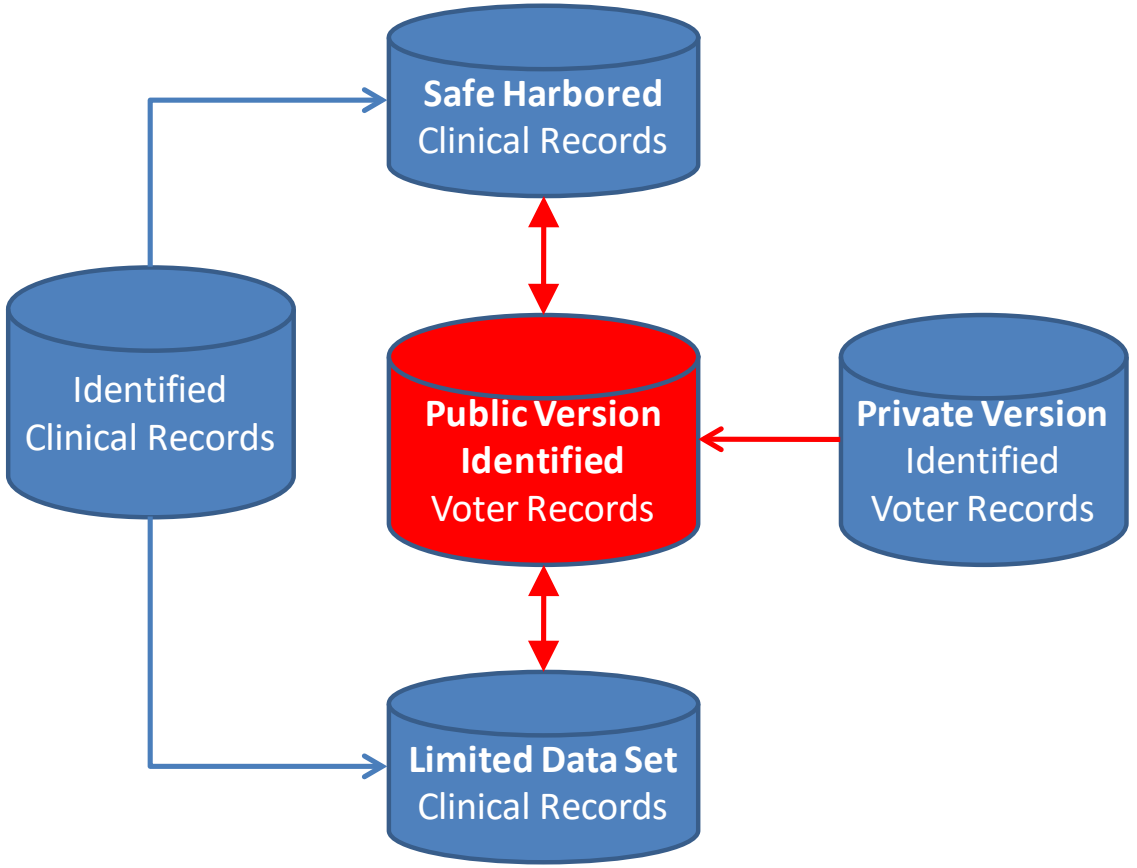
Central Dogma of Re-identification



Going to the Source

- They polled all U.S. states for what voter information is collected & shared

- What fields are shared?
- Who has access?
- Who can use it?
- What's the cost?



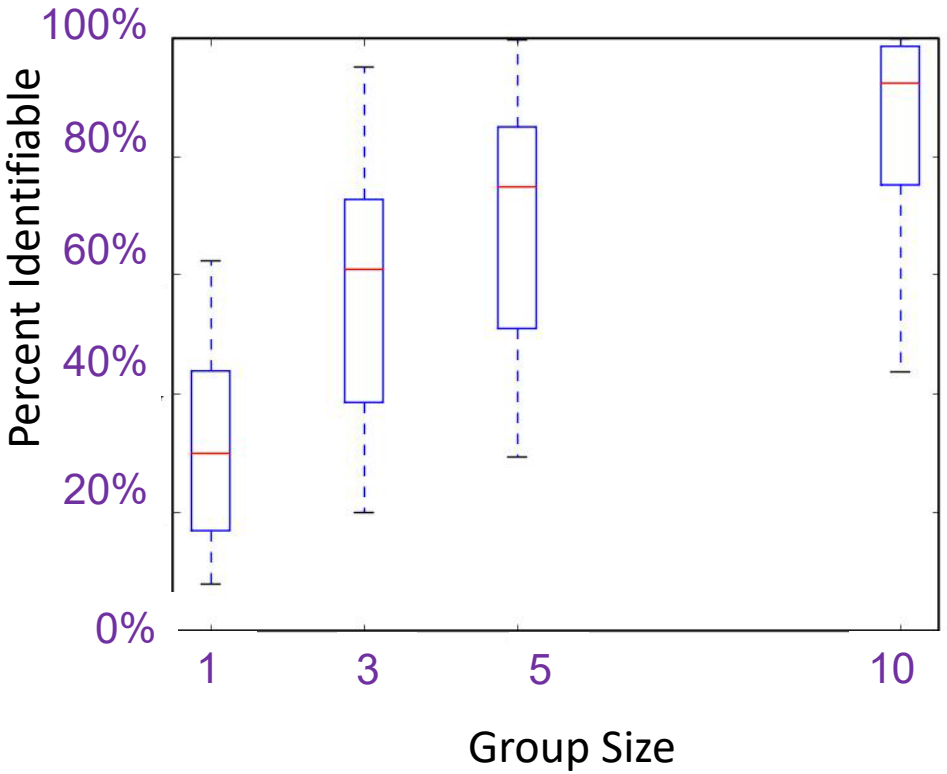
The Availability of Demographics Varies…

	IL	MN	TN	WA	WI
WHO	Registered Political Committees (ANYONE – In Person)	MN Voters	Anyone	Anyone	Anyone
Format	Disk	Disk	Disk	Disk	Disk
Cost	$500	$46; “use ONLY for elections, political activities, or law enforcement”	$2500	$30	$12,500
Name	●	●	●	●	●
Address	●	●	●	●	●
Date of Birth	●	○	●	●	
Sex	●		●	●	
Race			●		
Phone Number	●	●			

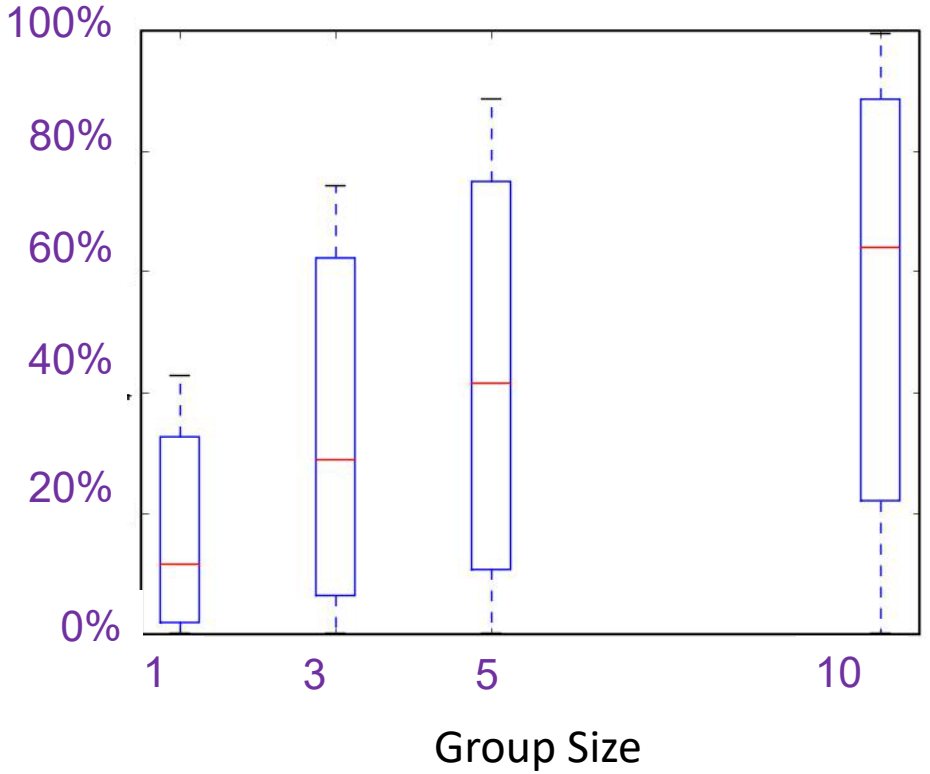
K. Benitez and B. Malin. Evaluating re-identification risks with respect to the HIPAA Privacy Rule. JAMIA. 2010; 17: 169-177.

Identifiability Changes!

Limited Data Set

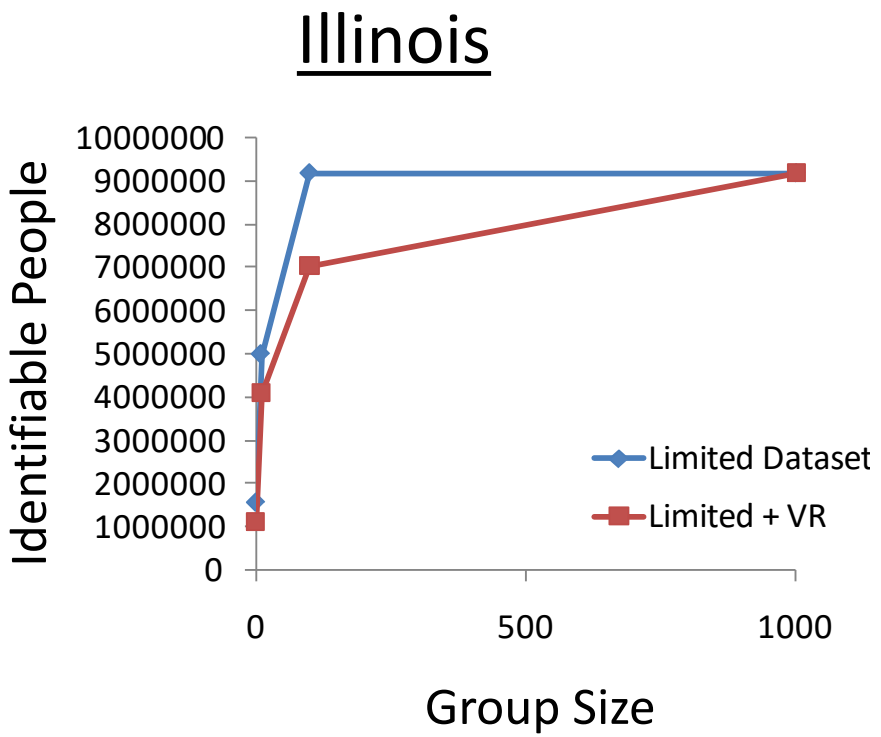


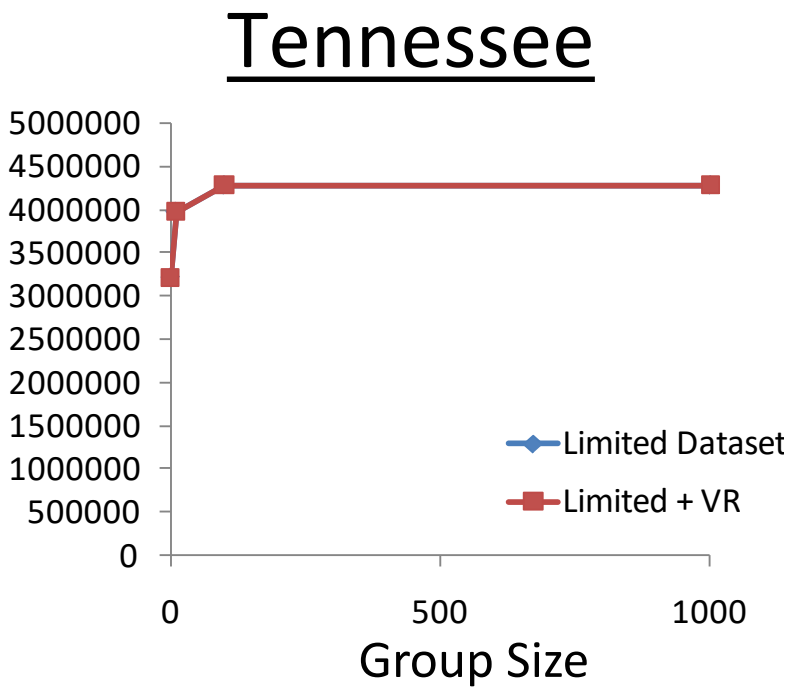
Limited Data Set ←→ Voter Reg.



K. Benitez and B. Malin. Evaluating re-identification risks with respect to the HIPAA Privacy Rule. JAMIA. 2010; 17: 169-177.

Worst Case vs. Reality





And What About Cost?

U.S. State	Limited Dataset		Safe Harbor	
	At Risk	Cost per Re-id	At Risk	Cost per Re-id
Virginia	3159764	$0	221	$0
South Carolina	2231973	$0	1386	$0
Wisconsin	72	$174	2	$6,250
West Virginia	55	$309	1	$17,000

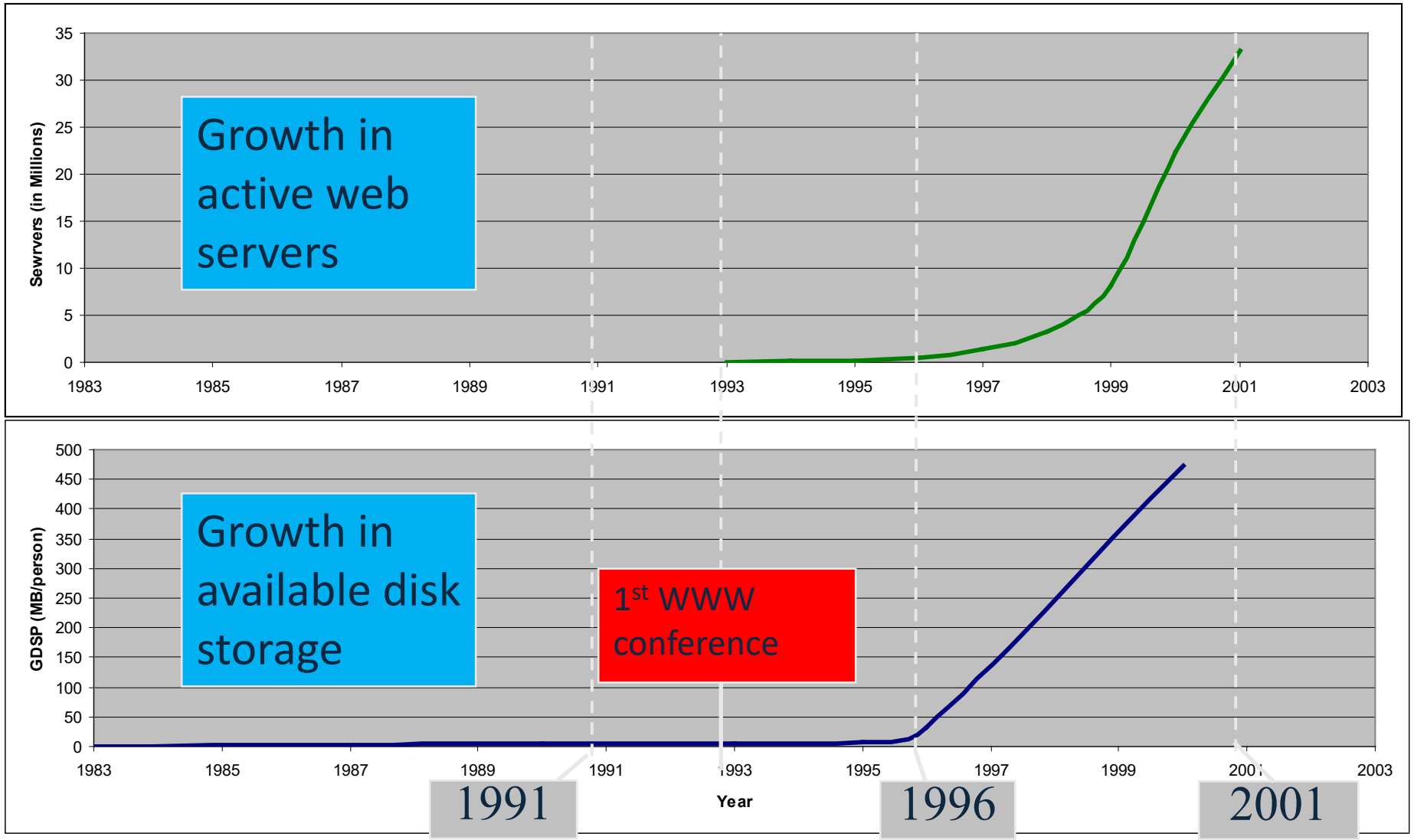
A “Real” Study Reported in Late 2010!

- Challenge issued by U.S. Dept. Health & Human Services
- Provided ~15,000 Safe Harbor records to re-id team @ U. Chicago on self identified “minority” ethnicity
- Team purchased public records from commercial broker
- Could correctly identified 2 people
- Details at http://www.ehcca.com/presentations/HIPAAWest4/lafky_2.pdf

Data Availability

- Information Generation
- Models of Availability
- Some Resources
- A Look at Voter Registration
- Curse of the SSN

Information Explosion



L. Sweeney. Information explosion. In L. Zayatz, et al. (eds) Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies . Urban Institute, Washington, DC, 2001

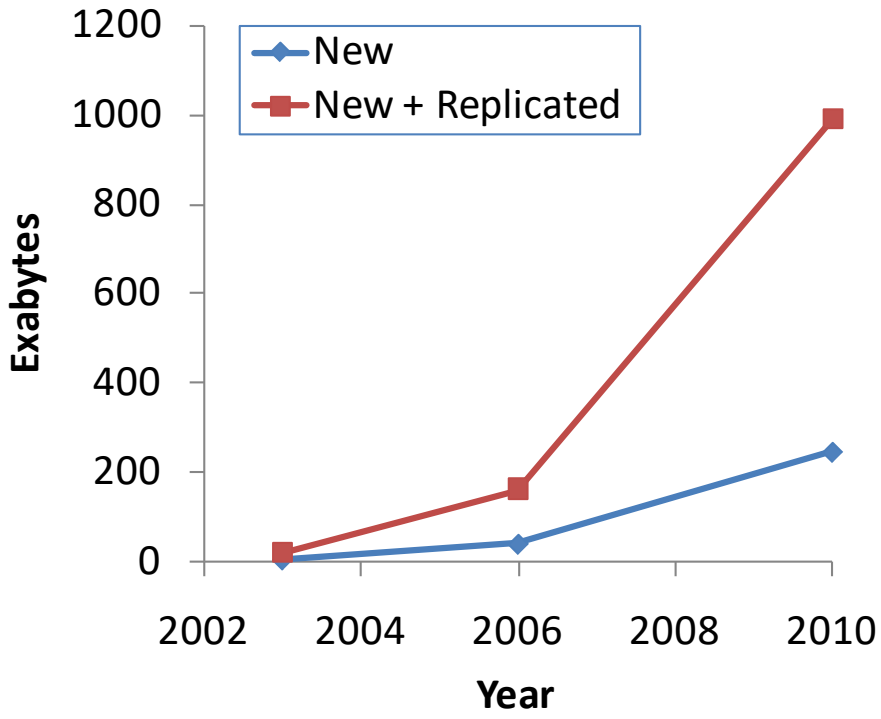
Newer Data

- Exabyte = 1 billion gigabytes

  * Includes analog data
    - radio communications, paper memos, etc.

  ** Includes new and replicated data
    - Some estimates put original information closer to 40 exabytes

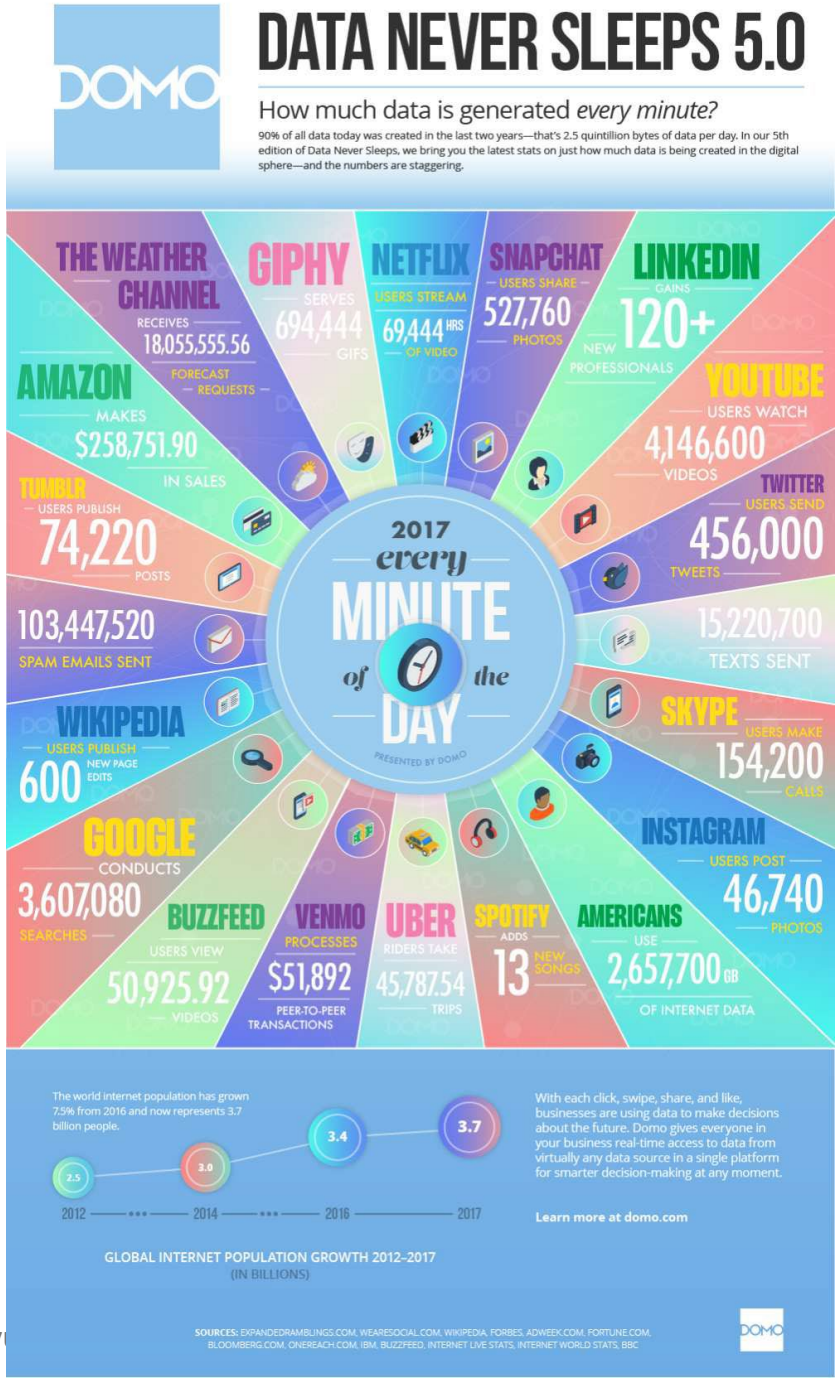
	New Data Generated in the "World"	Source
2003	5 exabytes*	UC Berkeley
2006	161 exabytes**	IDC



More recent numbers

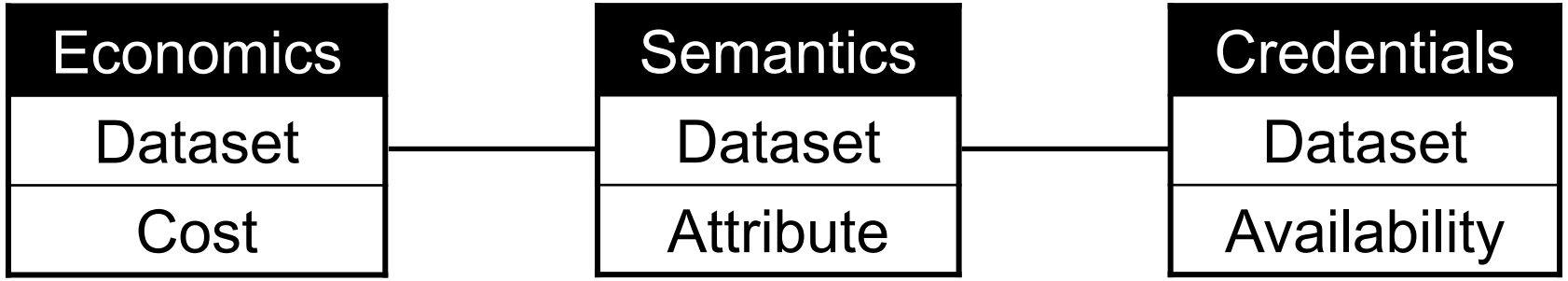
- On average, the US alone is generating 2,657,700 GB of *Internet data* every minute

- https://www.domo.com/learn/data-never-sleeps-5



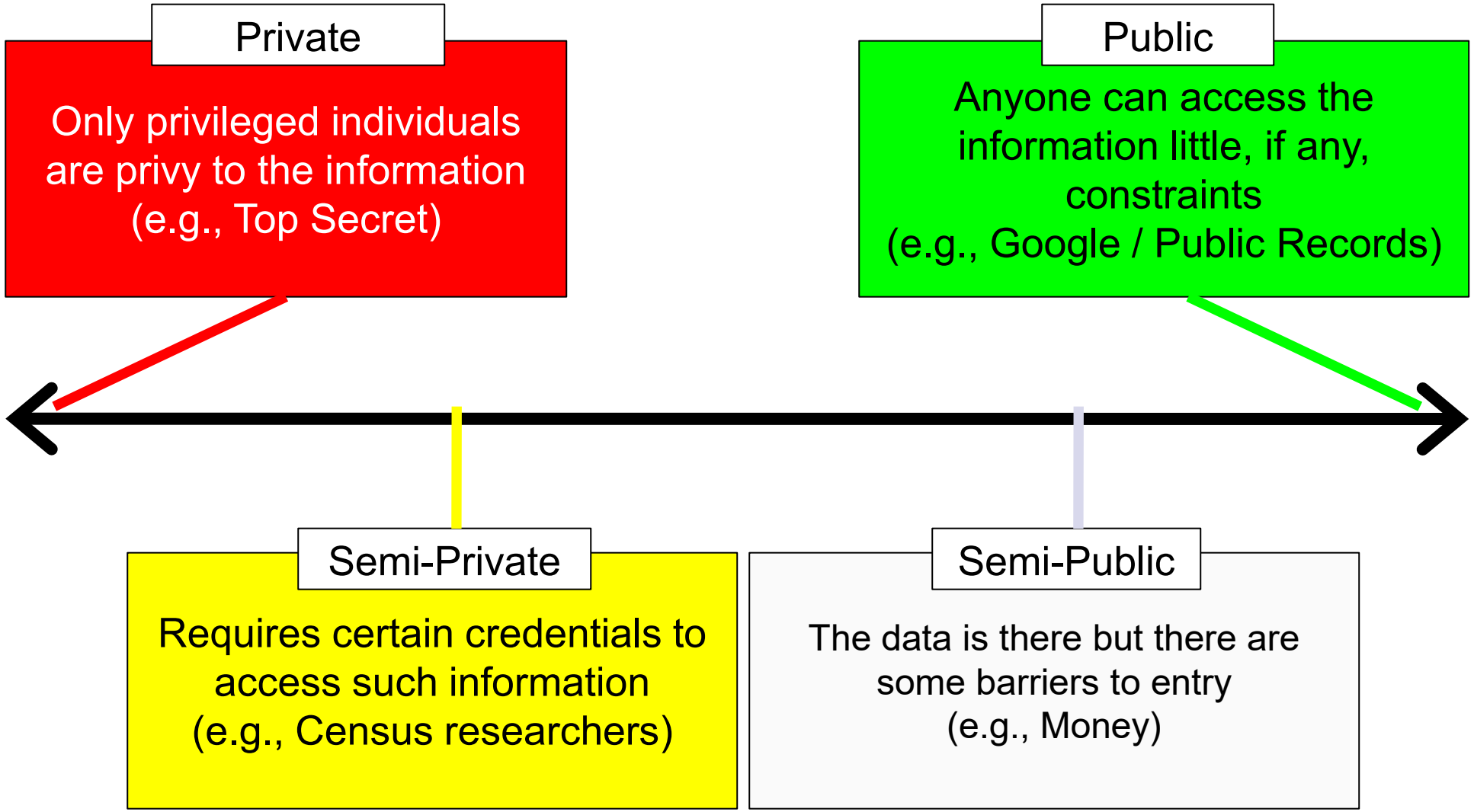
Accessibility

- Characterization of datasets / data
- Meta-information

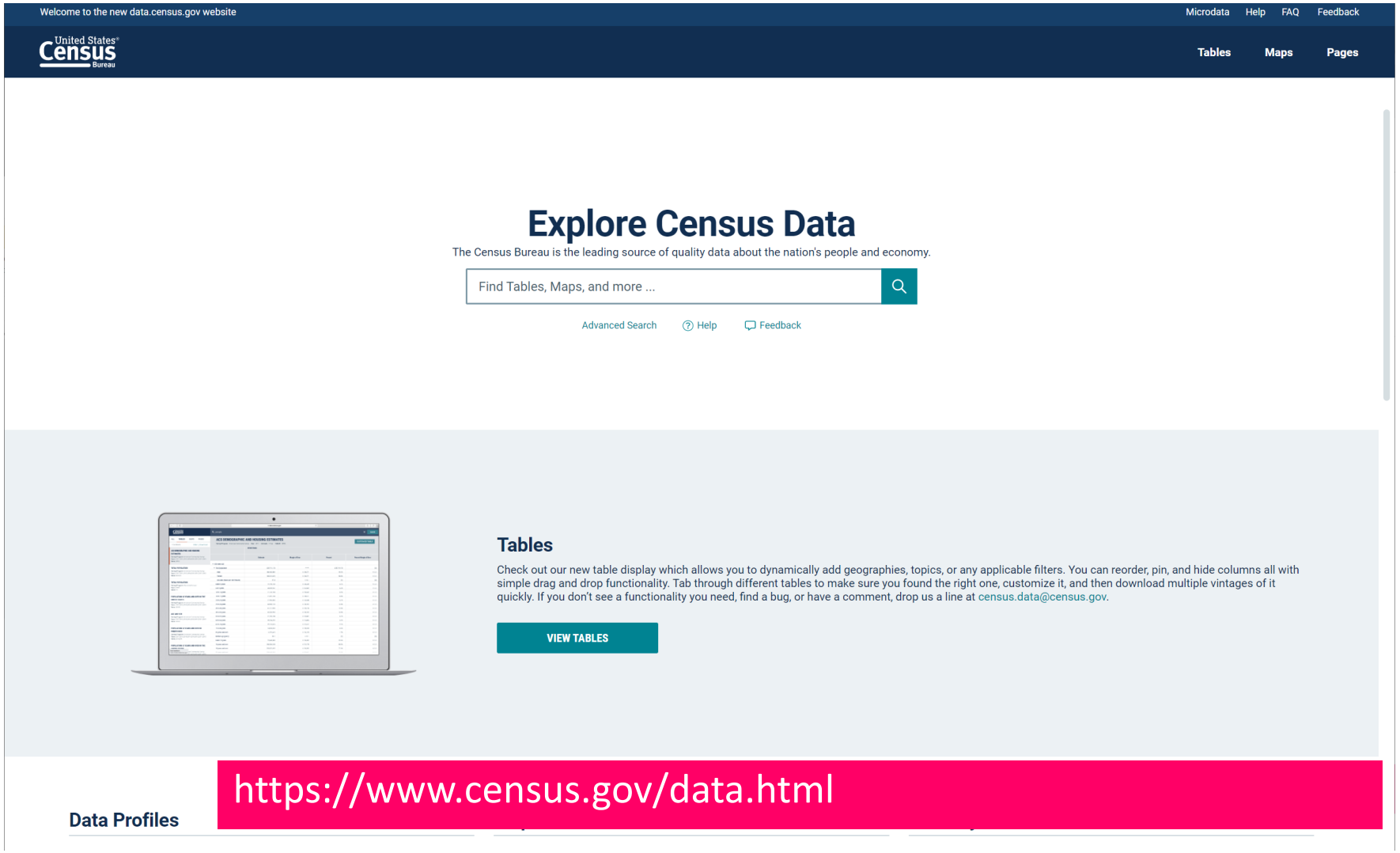


- Cost: Price per record or cost per dataset?
- Attribute: Type of data (e.g., name, birthdate, profession)
- Availability = Credentials needed to access the dataset

Availability

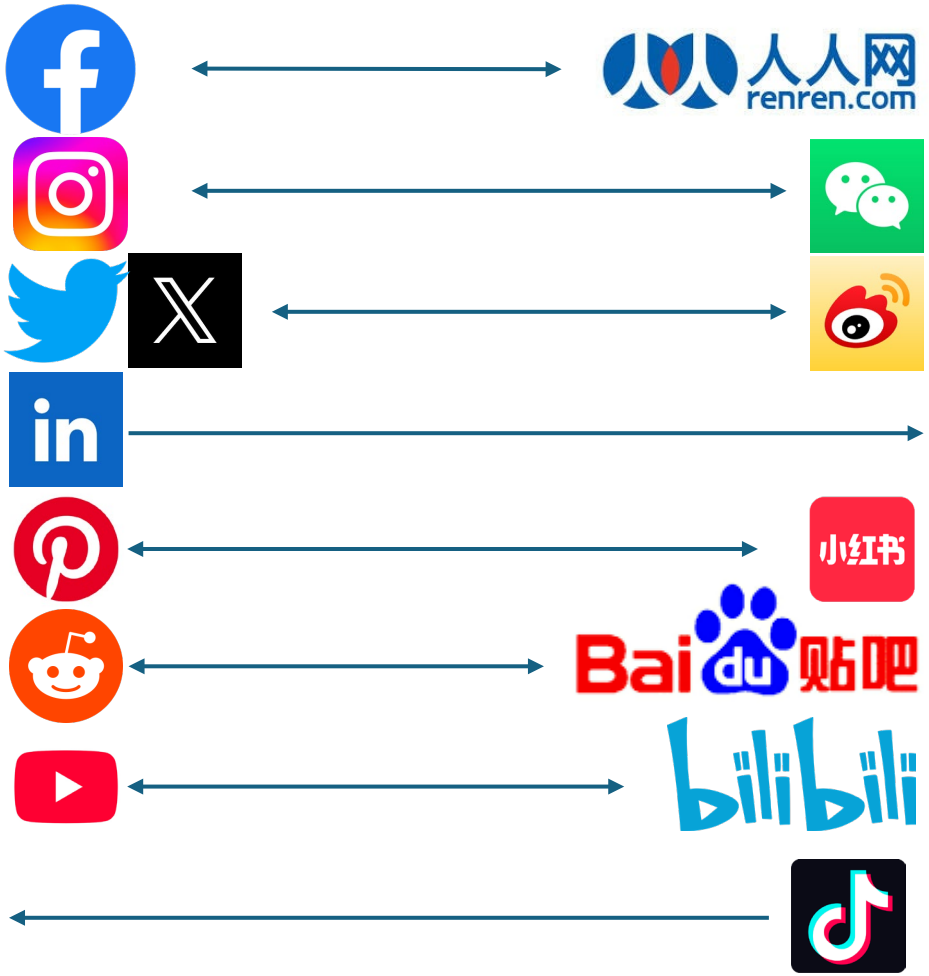


US Census


Welcome to the new data.census.gov website
Microdata
Help
FAQ
Feedback
United States Census Bureau
Tables
Maps
Pages
Explore Census Data
The Census Bureau is the leading source of quality data about the nation's people and economy.
Find Tables, Maps, and more ...
Advanced Search
Help
Feedback
Tables
Check out our new table display which allows you to dynamically add geographies, topics, or any applicable filters. You can reorder, pin, and hide columns all with simple drag and drop functionality. Tab through different tables to make sure you found the right one, customize it, and then download multiple vintages of it quickly. If you don't see a functionality you need, find a bug, or have a comment, drop us a line at census.data@census.gov.
VIEW TABLES
Data Profiles
https://www.census.gov/data.html

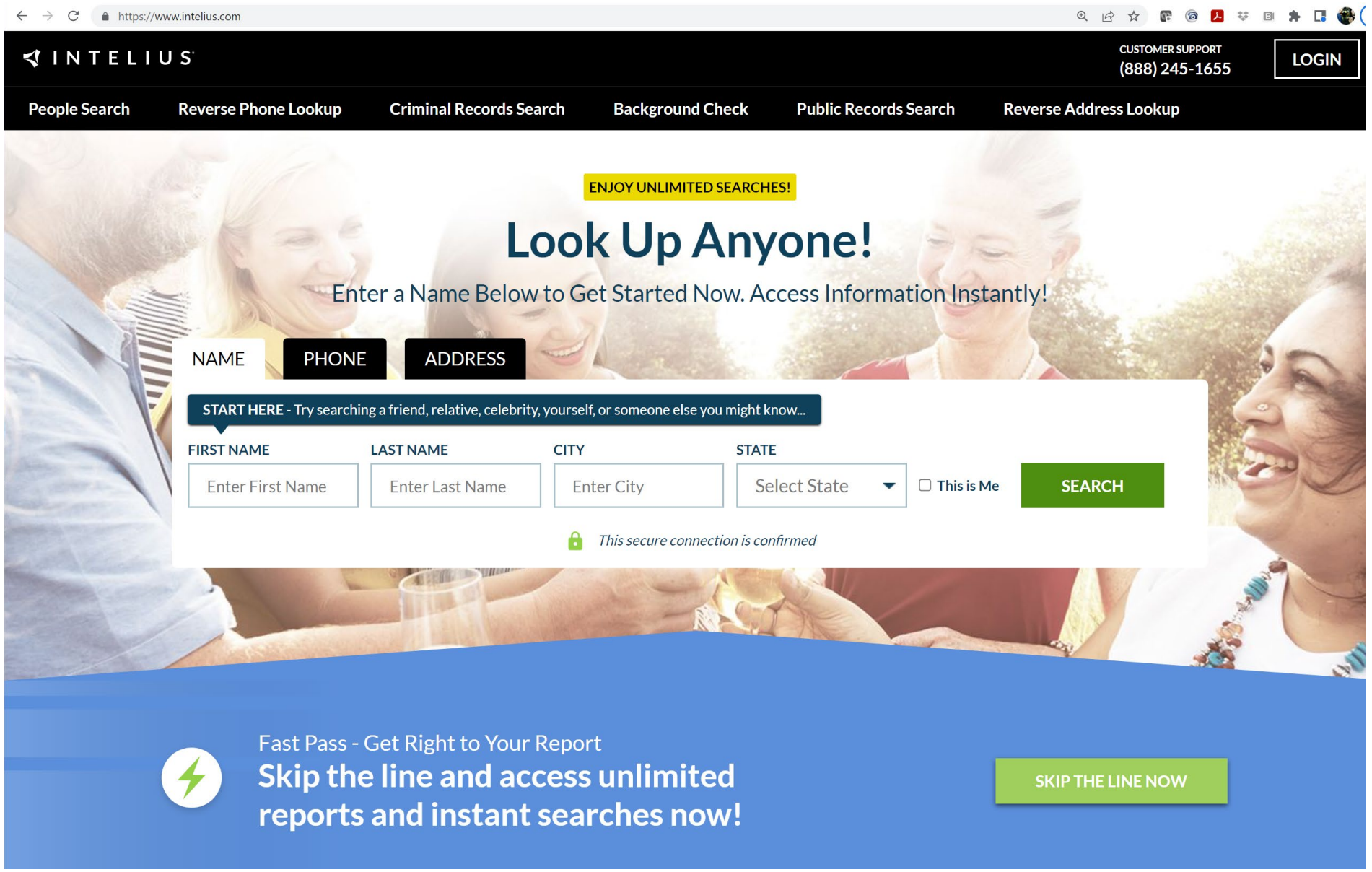
Social Media Platforms (US vs. China)

- Facebook
- Instagram
- Twitter/X
- LinkedIn
- Pinterest
- Reddit
- YouTube
- TikTok



- 人人网
- 微信（朋友圈）
- 新浪微博
- 领英
- 小红书
- 百度贴吧
- Bilibili
- 抖音

Intelius.com



Property Assessments

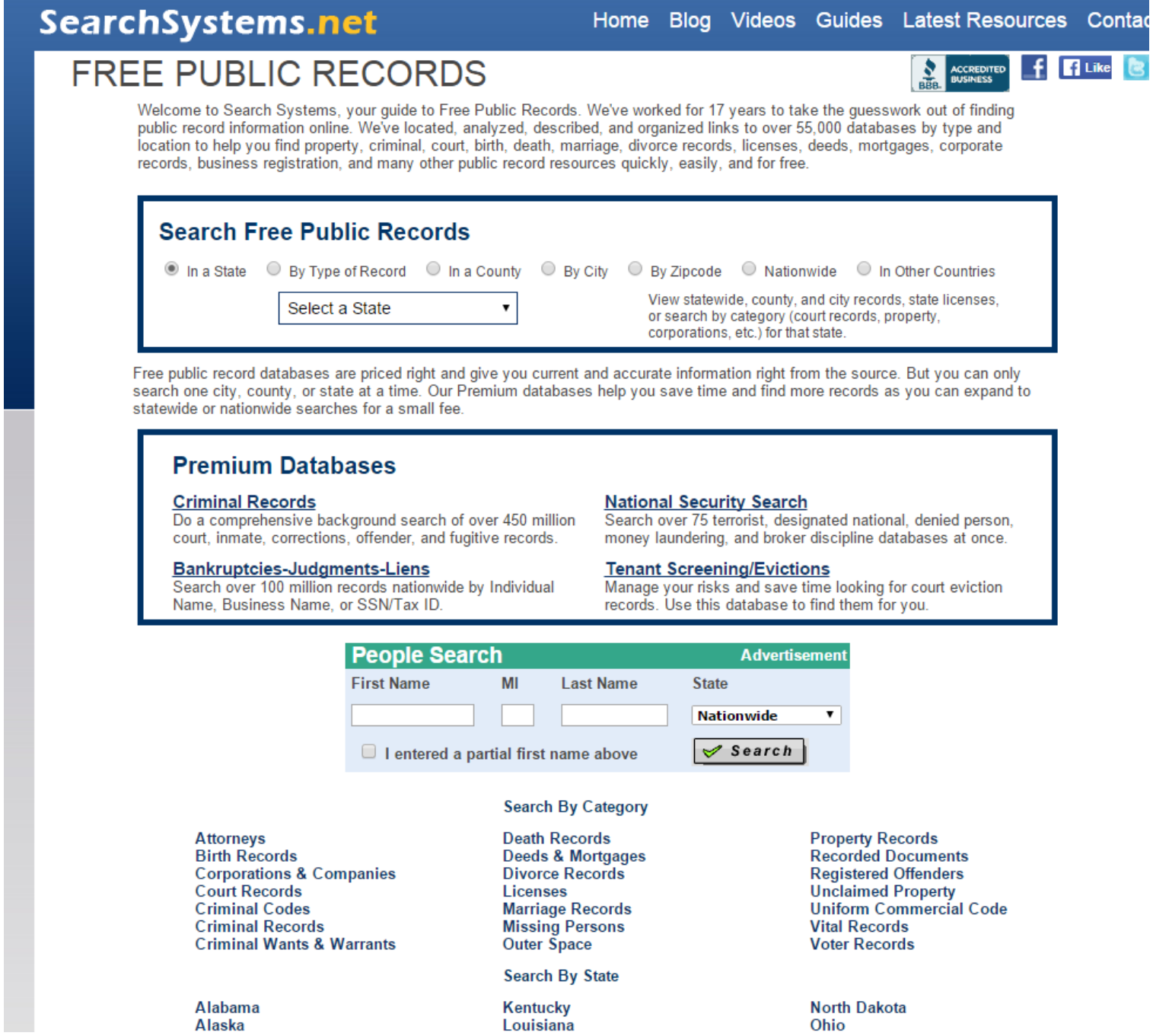
- Tennessee
  - http://www.assessment.cot.tn.gov/RE_Assessment/
- Davidson County
  - http://www.padctn.org/real-property-search/
  - Search by {Owner, Parcel, Street Address}

- Imagine combining with Google Maps (http://maps.google.com) or Zillow (http://www.zillow.com)

Reverse Lookups

- Phone
  - http://www.anywho.com/reverse-lookup
  - http://www.reversephonedirectory.com/

- DNS
  - http://remote.12dt.com/
  - http://psacake.com/web/eg.asp
  - http://www.dnsstuff.com/


SearchSystems.net
Home Blog Videos Guides Latest Resources Contac
FREE PUBLIC RECORDS
ACCREDITED BUSINESS
Like
Welcome to Search Systems, your guide to Free Public Records. We've worked for 17 years to take the guesswork out of finding public record information online. We've located, analyzed, described, and organized links to over 55,000 databases by type and location to help you find property, criminal, court, birth, death, marriage, divorce records, licenses, deeds, mortgages, corporate records, business registration, and many other public record resources quickly, easily, and for free.
Search Free Public Records
In a State
By Type of Record
In a County
By City
By Zipcode
Nationwide
In Other Countries
Select a State
View statewide, county, and city records, state licenses, or search by category (court records, property, corporations, etc.) for that state.
Free public record databases are priced right and give you current and accurate information right from the source. But you can only search one city, county, or state at a time. Our Premium databases help you save time and find more records as you can expand to statewide or nationwide searches for a small fee.
Premium Databases
Criminal Records
Do a comprehensive background search of over 450 million court, inmate, corrections, offender, and fugitive records.
National Security Search
Search over 75 terrorist, designated national, denied person, money laundering, and broker discipline databases at once.
Bankruptcies-Judgments-Liens
Search over 100 million records nationwide by Individual Name, Business Name, or SSN/Tax ID.
Tenant Screening/Evictions
Manage your risks and save time looking for court eviction records. Use this database to find them for you.
People Search
Advertisement
First Name
MI
Last Name
State
Nationwide
I entered a partial first name above
Search
Search By Category
Attorneys
Birth Records
Corporations & Companies
Court Records
Criminal Codes
Criminal Records
Criminal Wants & Warrants
Death Records
Deeds & Mortgages
Divorce Records
Licenses
Marriage Records
Missing Persons
Outer Space
Property Records
Recorded Documents
Registered Offenders
Unclaimed Property
Uniform Commercial Code
Vital Records
Voter Records
Search By State
Alabama
Alaska
Kentucky
Louisiana
North Dakota
Ohio

Collections on Everything

- Bankruptcy
- Birth
- Criminal
- Death
- Divorce
- DNS
- Employment
- Financial (e.g., donations)
- Marriage
- Military
- Residential
- Social Security
- Phone
- Voting
- …

Brokers are Real

Dear, Richards Jerome Warren sr

Complete Physician Database:

785K doctors
* First name * Last name * Title * Specialty * Address (city, state, zip, county)
* Medical School Attended * Location of Residency Training * Phone
* Fax * Email * Website * Primary Specialty * Secondary Specialty * Graduation Year
* Major Activity * Hospital * Group Practice Record Count By State

Here is the package deal we're running for December

Price Just Lowered - $394.50

***Get the data below as a gift when you order the MD list above ***

Chiropractor Database: (reg.249.00)- December OFFER 199.00)
108,421 Total Records *3,414 Emails *6,553 Fax Numbers

Pharmacy Database: (reg.249.00)- December OFFER 199.00)
143,552 Total Records *1,193 Emails *1,700 Fax Numbers *Sortable fields Available

Veterinarian Database: (reg.199.00)- December OFFER 149.00)
78,986 Total Records *1,438 Emails *1,050 Fax Numbers *Sortable fields Available

Alternative Medicine Database: (reg.599.00)- December OFFER 399.00) There are
1,141,602 total records with 72,320 emails and 38.935 fax numbers.

reply to: rubenrowley@hotmail.com

Brokers are Big Business

Take a look at

- ConcertAI
- HealthVerity,
- IQVIA
- Komodo Health,
- Loopback Analytics
- Purple Health
- …

- Datavant Switchboard

IBM sells Watson Health assets to investment firm Francisco Partners

By **Heather Landi** • Jan 21, 2022 08:55am

Artificial Intelligence | Data Analytics | mergers and acquisitions | Oncology

IBM launched Watson Health in April 2015 to offer cloud-based access to the company's Watson supercomputer for analyzing healthcare data. According to reports, IBM made a series of acquisitions to build Watson Health to the tune of $4 billion. (AndreyPopov/GettyImages)

IBM has reached a deal to sell the healthcare data and analytics assets from its Watson Health business to investment firm Francisco Partners, the companies announced Friday.

The companies did not disclose financial terms of the deal, but previous reports said IBM was seeking a price in the $1 billion range, Axios reported, citing people familiar

BUSINESS INSIDER Subscribe

TECH

Here's what potential 23andMe buyers could do with your genetic data

By **Lakshmi Varanasi** and **Lauren Edmonds**

Getty Images; Alyssa Powell/BI

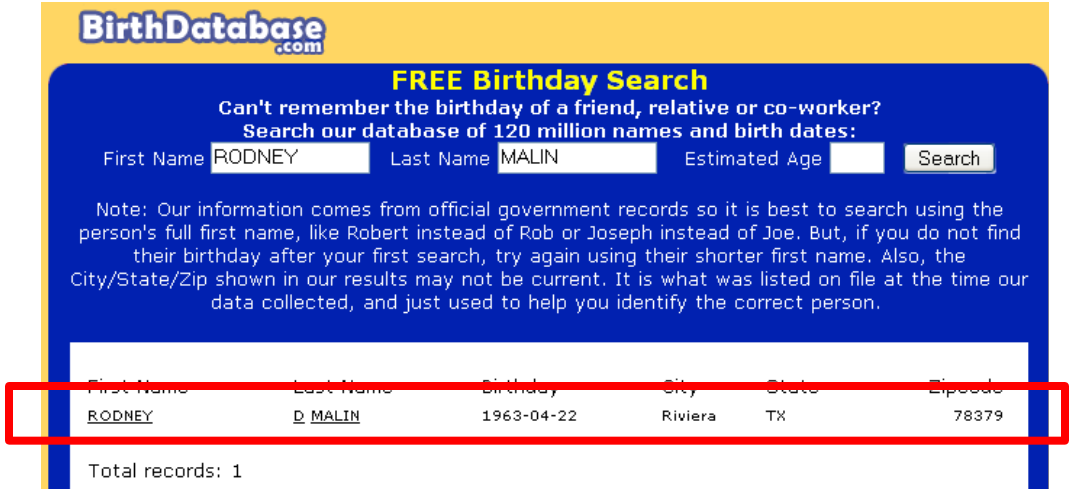
Apr 14, 2025, 1:26 AM GMT+8 Share Save

- **23andMe filed for bankruptcy amid financial struggles and fallout after a data breach.**
- **Now, potential buyers and investors are weighing the value of its DNA database.**
- **One of the world's largest troves of genetic information, its sale is raising new privacy concerns.**

Late last month, genetic testing giant 23andMe filed for bankruptcy.

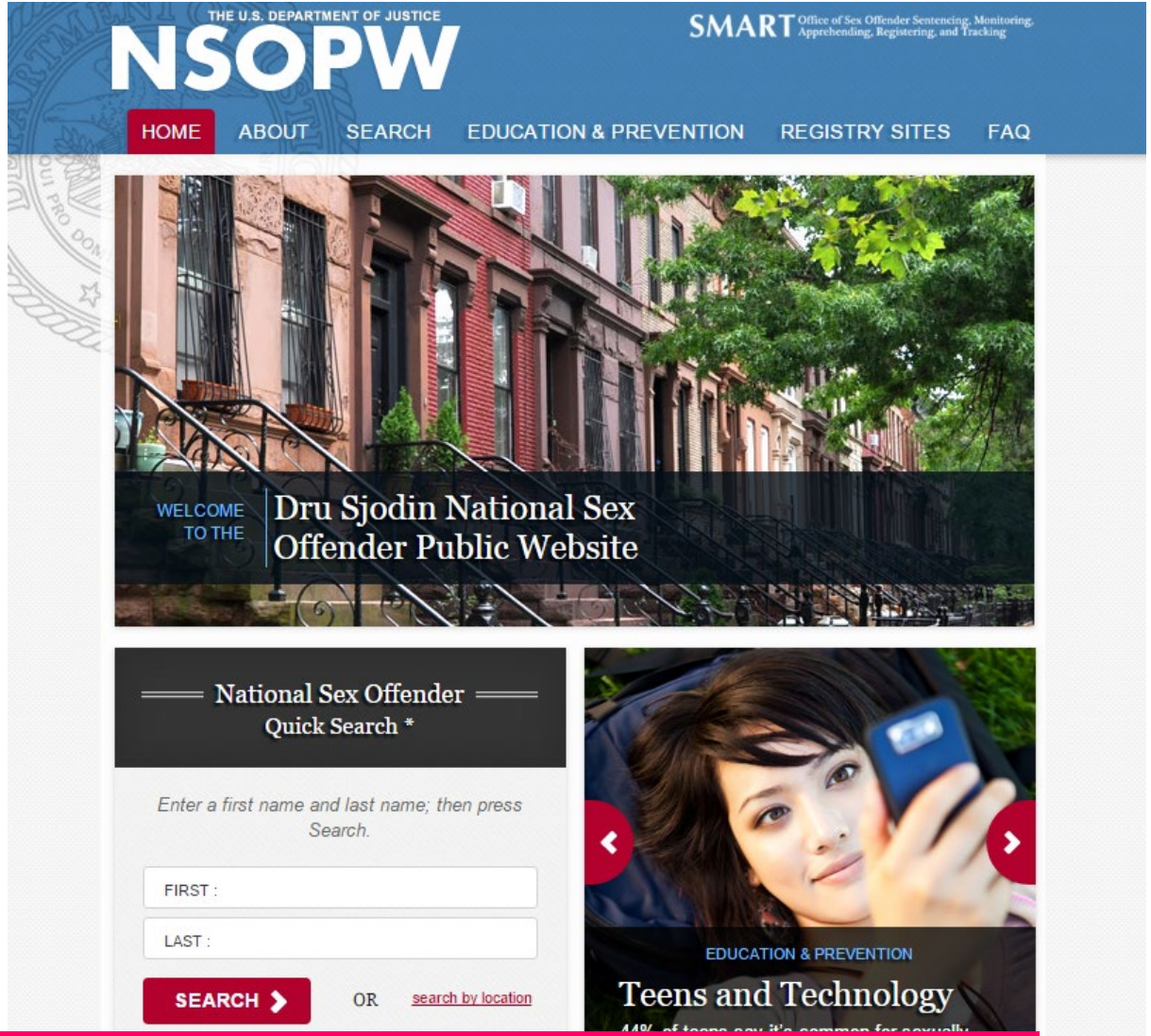
Combining Databases?

- How do you integrate these databases?
  - Do you trust names?
  - Do you trust phone numbers?
  - How much information would you need until you're confident of a match?



		NAME	AGE	PREVIOUS CITIES	DOB	PHONE	ADDRESS	RELATIVES
11	View Details	THOMAS D MALIN	37	MADISON, TN SPRINGFIELD, TN COOKEVILLE, TN NASHVILLE, TN JOELTON, TN	✔	✔	✔	THOMAS D MALIN MICHELLE D MALIN TOM D MALIN
12	View Details	DARYL LEE MALIN	54	LENOIR CITY, TN	✔	✔	✔	DARYL LEE MALIN MARCIA M MALIN MIC MALIN
13	View Details	HOLLIS H MALIN	95	KNOXVILLE, TN	✔	✔	✔	HOLLIS H MALIN B MALIN
14	View Details	THOMAS D MALIN	57	GOODLETTSVILLE, TN	✔	✔	✔	THOMAS D MALIN MARY J MALIN
15	View Details	RODNEY MALIN	-	COOKEVILLE, TN MC MINNVILLE, TN MORRISTOWN, TN FRANKLIN, TN	-	✔	✔	RODNEY MALIN CHARLES RODNEY MALIN STEPHANY L MALIN SHERRY MALIN
16	View Details	ANNA MALIN MALIN	42	MEMPHIS, TN PADUCAH, KY	✔	✔	✔	ANNA MALIN MALIN
17	View Details	DEBORAH D MALIN	50	GERMANTOWN, TN MEMPHIS, TN	✔	✔	✔	DEBORAH D MALIN

Registries



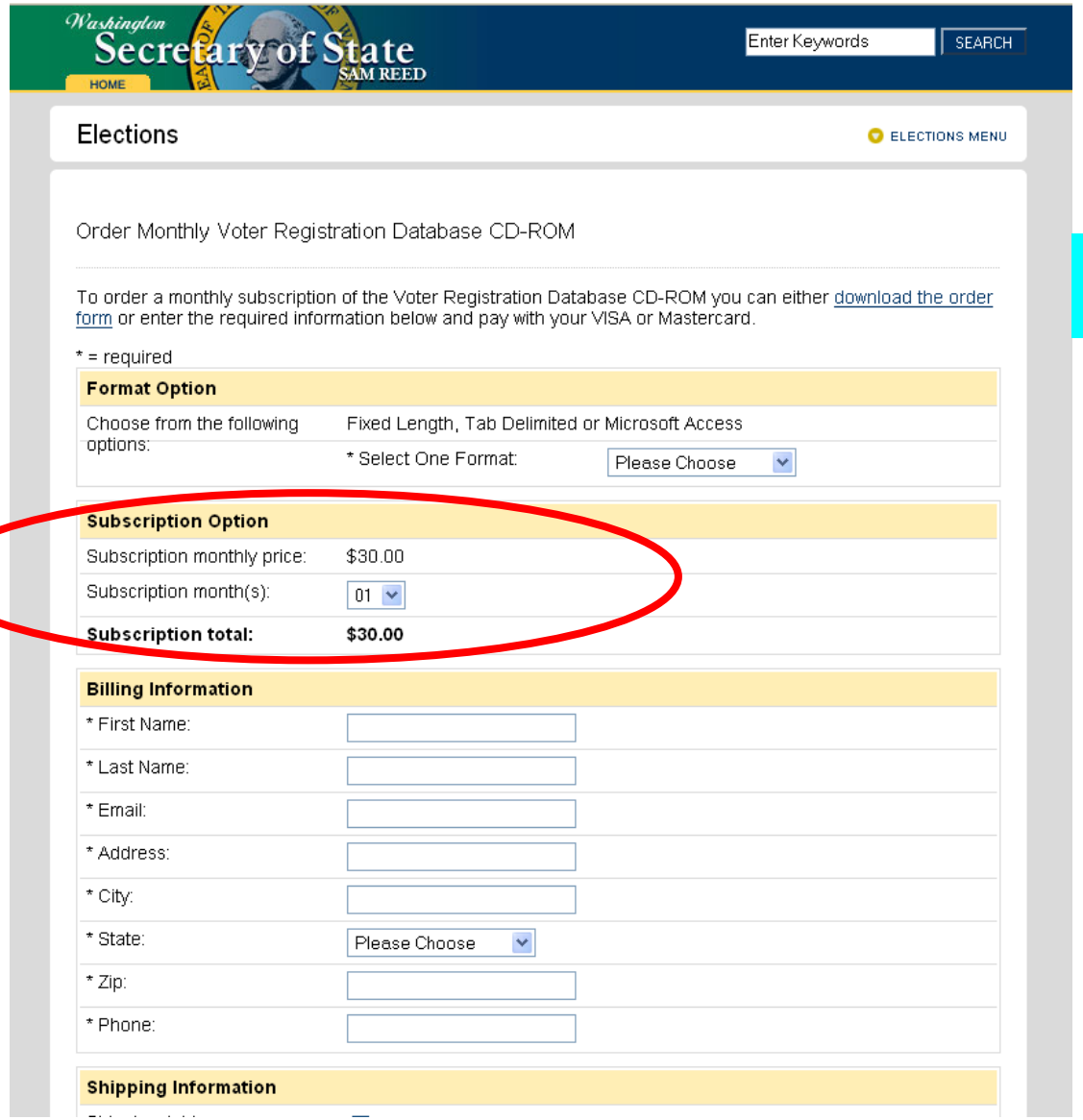
- National Sex Offender Registry
  - https://www.nsopw.gov/

Remember the Voter Database?

- Public Information Sharing
- Example: Washington
  - If you are a voter, your name, address, political jurisdiction, gender, date of birth, voting record, date of registration, and registration number are public information under state law.
    (RCW 29A.08.710)
- This is public record by state law and does not violate security or privacy policy, however

Behind Closed Doors

- More than public information for registration of voters
- Goal: Enable sharing of information with other states securely and **accurately** in fulfillment of HAVA (the Help America Vote Act)
- Example: Pennsylvania
  - "Statewide Uniform Registry of Electors", county election officials have direct access to the centralized statewide database
  - The state uses "identifying number, name, and date of birth" for linking to motor vehicle and/or Social Security records
    - Use a hybrid match: the number and first two characters of the last name must match exactly, with discretion left to the county commission to determine if the rest of the record is a match
    - Currently uses the AAMVA (American Association of Motor Vehicles Administrators) criteria to match information with SSN digits: exact match of the last four digits of Social Security Number, first name, last name, month of birth, and year of birth.


Washington
Secretary of State
SAM REED
HOME
Enter Keywords
SEARCH
Elections
ELECTIONS MENU
Order Monthly Voter Registration Database CD-ROM
To order a monthly subscription of the Voter Registration Database CD-ROM you can either download the order form or enter the required information below and pay with your VISA or Mastercard.
* = required
Format Option
Choose from the following options:
Fixed Length, Tab Delimited or Microsoft Access
* Select One Format:
Please Choose
Subscription Option
Subscription monthly price:
$30.00
Subscription month(s):
01
Subscription total:
$30.00
Billing Information
* First Name:
* Last Name:
* Email:
* Address:
* City:
* State:
Please Choose
* Zip:
* Phone:
Shipping Information

Washington

http://dl.ncsbe.gov/index.html

Index of /data/

Name	Size	Date Modified
[parent directory]		
idrequired/		5/30/13, 8:04:00 AM
ncvhis1.zip	3.1 MB	1/30/16, 5:16:00 AM
ncvhis10.zip	3.0 MB	1/30/16, 5:21:00 AM
ncvhis100.zip	512 kB	1/30/16, 6:21:00 AM
ncvhis11.zip	7.6 MB	1/30/16, 5:22:00 AM
ncvhis12.zip	2.0 MB	1/30/16, 5:23:00 AM
ncvhis13.zip	3.8 MB	1/30/16, 5:24:00 AM
ncvhis14.zip	1.7 MB	1/30/16, 5:25:00 AM
ncvhis15.zip	177 kB	1/30/16, 5:25:00 AM
ncvhis16.zip	1.8 MB	1/30/16, 5:26:00 AM
ncvhis17.zip	426 kB	1/30/16, 5:26:00 AM
ncvhis18.zip	3.5 MB	1/30/16, 5:27:00 AM
ncvhis19.zip	1.9 MB	1/30/16, 5:27:00 AM
ncvhis2.zip	840 kB	1/30/16, 5:17:00 AM
ncvhis20.zip	607 kB	1/30/16, 5:28:00 AM
ncvhis21.zip	305 kB	1/30/16, 5:28:00 AM
ncvhis22.zip	325 kB	1/30/16, 5:29:00 AM
ncvhis23.zip	2.4 MB	1/30/16, 5:29:00 AM
ncvhis24.zip	1.2 MB	1/30/16, 5:30:00 AM
ncvhis25.zip	2.2 MB	1/30/16, 5:30:00 AM
ncvhis26.zip	7.3 MB	1/30/16, 5:32:00 AM
ncvhis27.zip	452 kB	1/30/16, 5:32:00 AM
ncvhis28.zip	917 kB	1/30/16, 5:33:00 AM

North Carolina

Policy & Usage Restrictions

RCW 29A.08.740

Violations of restricted use of registered voter data - Penalties - Liabilities. (Effective January 1, 2006.)

(1) **Any person who uses registered voter data furnished under RCW 29A.08.720 for the purpose of mailing or delivering any advertisement or offer for any property, establishment, organization, product, or service or for the purpose of mailing or delivering any solicitation** for money, services, or anything of value is guilty of a **class C felony** punishable by imprisonment in a state correctional facility for a period of not more than five years or a fine of not more than ten thousand dollars or both such fine and imprisonment, and is liable to each person provided such advertisement or solicitation, without the person's consent, for the nuisance value of such person having to dispose of it, which value is herein established at five dollars for each item mailed or delivered to the person's residence. …

(2) Each person furnished data under RCW 29A.08.720 shall take reasonable precautions designed to assure that the **data is not used for the purpose of mailing or delivering any advertisement or offer for any property, establishment, organization, product, or service or for the purpose of mailing or delivering any solicitation for money, services, or anything of value.** However, the data may be used for any political purpose. Where failure to exercise due care in carrying out this responsibility results in the data being used for such purposes, then such person is jointly and severally liable for damages under subsection (1) of this section along with any other person liable under subsection (1) of this section for the misuse of such data.

[2005 c 246 § 19. Prior: 2003 c 111 § 249; 2003 c 53 § 176; 1999 c 298 § 2; 1992 c 7 § 32; 1974 ex.s. c 127 § 3; 1973 1st ex.s. c 111 § 4. Formerly RCW 29.04.120.]

SSNs – Who Cares?

- The Social Security Number is one of, if not, ***the*** most overloaded numbers in the US
- It binds records on finances, insurance, education, death, taxes…
- Two HUGE problems: Fraud & Identity Theft

Year	Total # of Complaints	% of Complaints Reporting Amount Paid	Reported Amount Paid	Avg. Amount Paid	Median Amount Paid
2003	327,479	78%	$459M	$1.8k	$222
2004	406,193	76%	$567M	$1.8k	$267
2005	431,118	66%	$682M	$2.4k	$350
2016	3,000,000	51%	$744M	$1.1k	$450

- 63% Fraud: Internet Auction (12%), Foreign Money Offer (8%)
- 37% ID theft: Credit card (26%), Phone / Utilities (18%), Employment (12%), Government documents / benefits (9%), Loan (5%)

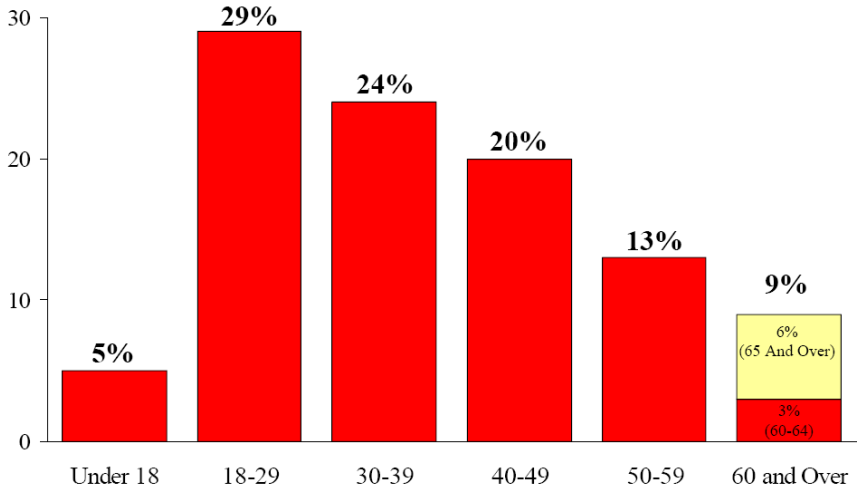
https://www.ftc.gov/system/files/documents/reports/consumer-sentinel-network-data-book-january-december-2016/csn_cy-2016_data_book.pdf

Location and Age Matter (2005)

- Highest per capita rate of identity theft (metropolitan areas)

Rank	Metropolitan Area	Complaints Per 100,000 People
1	Phoenix-Mesa-Scottsdale, AZ	178.3
2	Las Vegas-Paradise, NV	158.5
3	Riverside-San Bernardino-Ontario, CA	145.7
43	Nashville-Davidson--Murfreesboro, TN	63.6

- Rate of victimization by age range



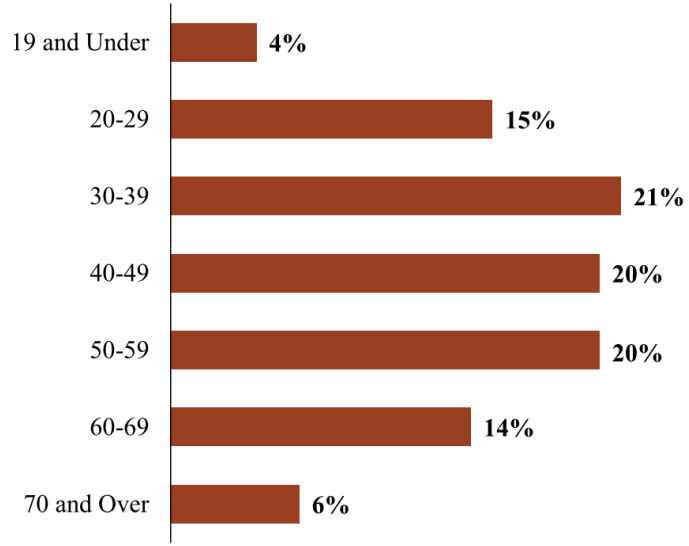
https://www.ftc.gov/system/files/documents/reports/consumer-sentinel-network-data-book-january-december-2016/csn_cy-2016_data_book.pdf

Location and Age Matter (2016)

- Highest per capita rate of identity theft (metropolitan areas)

Rank	Metropolitan Area	Complaints Per 100,000 Population[1]	Complaints
1	Ann Arbor, MI Metropolitan Statistical Area	336.0	1,206
2	Miami-Fort Lauderdale-West Palm Beach, FL Metropolitan Statistical Area	223.8	13,457
3	Detroit-Warren-Dearborn, MI Metropolitan Statistical Area	214.7	9,238
4	St. Louis, MO-IL Metropolitan Statistical Area	206.0	5,792
5	Naples-Immokalee-Marco Island, FL Metropolitan Statistical Area	188.1	672

- Rate of victimization by age range



https://www.ftc.gov/system/files/documents/reports/consumer-sentinel-network-data-book-january-december-2016/csn_cy-2016_data_book.pdf

SSNs

- Federal paternalism for “social insurance”
- Benefits based on payroll tax contributions →

  Federal Old-Age Benefits
- Issuance began in late 1936
- Issued by the Social Security Administration
  - Permanent residents
  - Temporary / working residents
- First number to John David Sweeney Jr. (Baltimore, MD)
- Lowest number issued: 001-01-0001

John David Sweeney

http://www.ssa.gov/history

Medicare

- http://my.medicare.gov/
- Medicare Identification Number (MIN) is usually SSN + an added letter
- Ex: 000-00-0000A
  - *A* = wage earner (primary)
  - If spouse becomes eligible for Medicare benefits through primary, they are assigned a *B*
  - Many valid suffixes
- MIN may be different than the SSN

Ferree Snafu

- 1938: Wallet manufacturer, E.H. Ferree, promoted new wallet
- Sample card was a real card
  - Hilda Schrader Whitcher, secretary of the company vice president
- Wallet sold by Woolworth department stores across the USA
- 1943: 5,755 people using the number
- SSA voided the number; issued Hilda new card
- > 40,000 people reported the Whitcher number as their own
- 1976: 40 people found using the number
- 1977: 12 “ “ “ “ “

- It’s known as “the Social Security Number issued by Woolworth”

SSN Assignment

- SSNs are almost a one-time shot
- You can get a new SSN only in extremely rare circumstances
- You must prove
  - Someone has stolen your number
  - Someone is using it illegally
  - The misuse is causing you serious harm

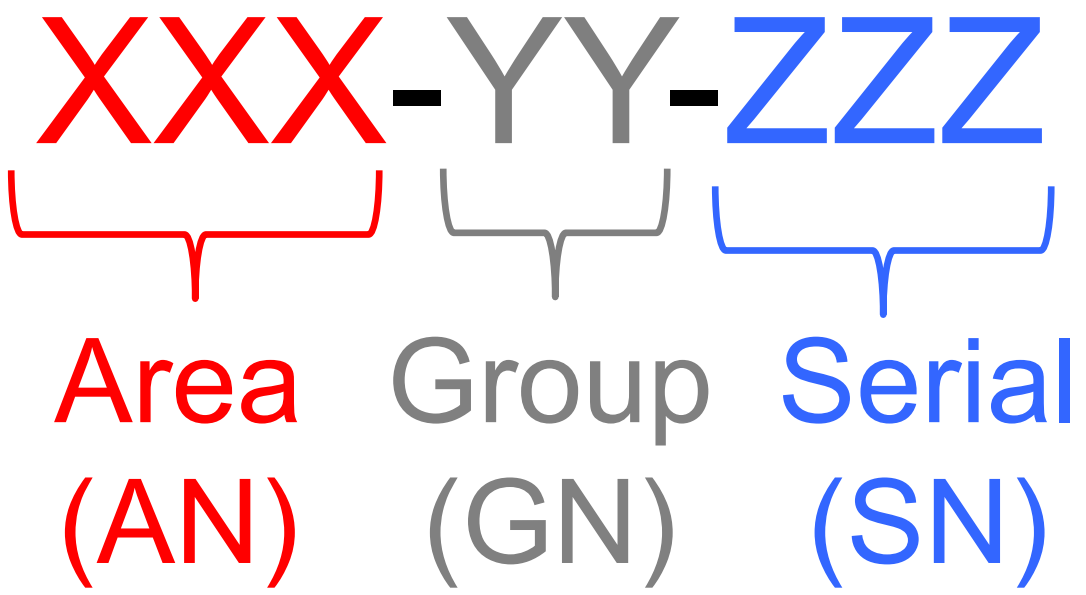
http://www.socialsecurity.gov/ssnumber/ss5doc.htm

Modern Times: Restricted Use

- Some state laws restrict SSN use, display, and transfer (e.g., CA in 2001)

- Michigan prohibits use of more than 4 consecutive digits of an SSN
  - Is that sufficient protection?

The SSN



Area Numbers: XXX

- Prior to 1972: represented the state from which a person applied for their social security card

- After 1972: based on zip code in the mailing address provided on the application form

Area Numbers: XXX

#	STATE	#	STATE	#	STATE	#	STATE
001-003	NH	232 237-246 681-690	NC	387-399	WI	627-645	TX
004-007	ME			400-407	KY	468-477	MN
008-009	VT			408-415 756-763	TN	478-485	IA
010-034	MA	247-251 654-658	SC			486-500	MO
035-039	RI			416-424	AL	501-502	ND
040-049	CT	252-260 667-675	GA	425-428 587-588 752-755	MI	503-504	SD
050-134	NY					505-508	NB
135-158	NJ	261-267 589-595 766-772	FL			509-515	KS
159-211	PA			429-432 676-679	AR	516-517	MT
212-220	MD					518-519	ID
221-222	DE	268-302	OH	433-439 659-665	LA	520	WY
223-231 691-699	VA	303-317	IN			521-524 650-653	CO
		318-361	IL	440-448	OK		
232-236	WVA	362-386	MI	449-467	TX		

Area Numbers: XXX

#	STATE	#	STATE
525, 585 648-649	NM	575-576 750-751	HI
526-527 600-601 764-765	AZ	577-579	DC
		580	Virgin Islands
		580-584 596-599	Puerto Rico
528-529 646-647	UT		
		586	Guam
530,680	NV	586	American Samoa
531-539	WA	586	Philipine Islands
540-544	OR	700-728	Railroad Board**
545-573 602-626	CA	729-733	Enumeration at Entry
574	AK		

- **Discontinued 7/1/63
- 000 will NEVER be a valid XXX number

Group Numbers: YY

- Range from 01-99
- They're not allocated consecutively!
  - Highest group issued as of 1/2/08
    - http://www.ssa.gov/employer/highgroup.txt

Order	Type	Range
1st	Odd	01 through 09
2nd	Even	10 through 98
3rd	Even	02 through 08
4th	Odd	11 through 99

  - Can also trace the allocation of group numbers over time:
    - http://www.ssa.gov/employer/ssnvhighgroup.htm

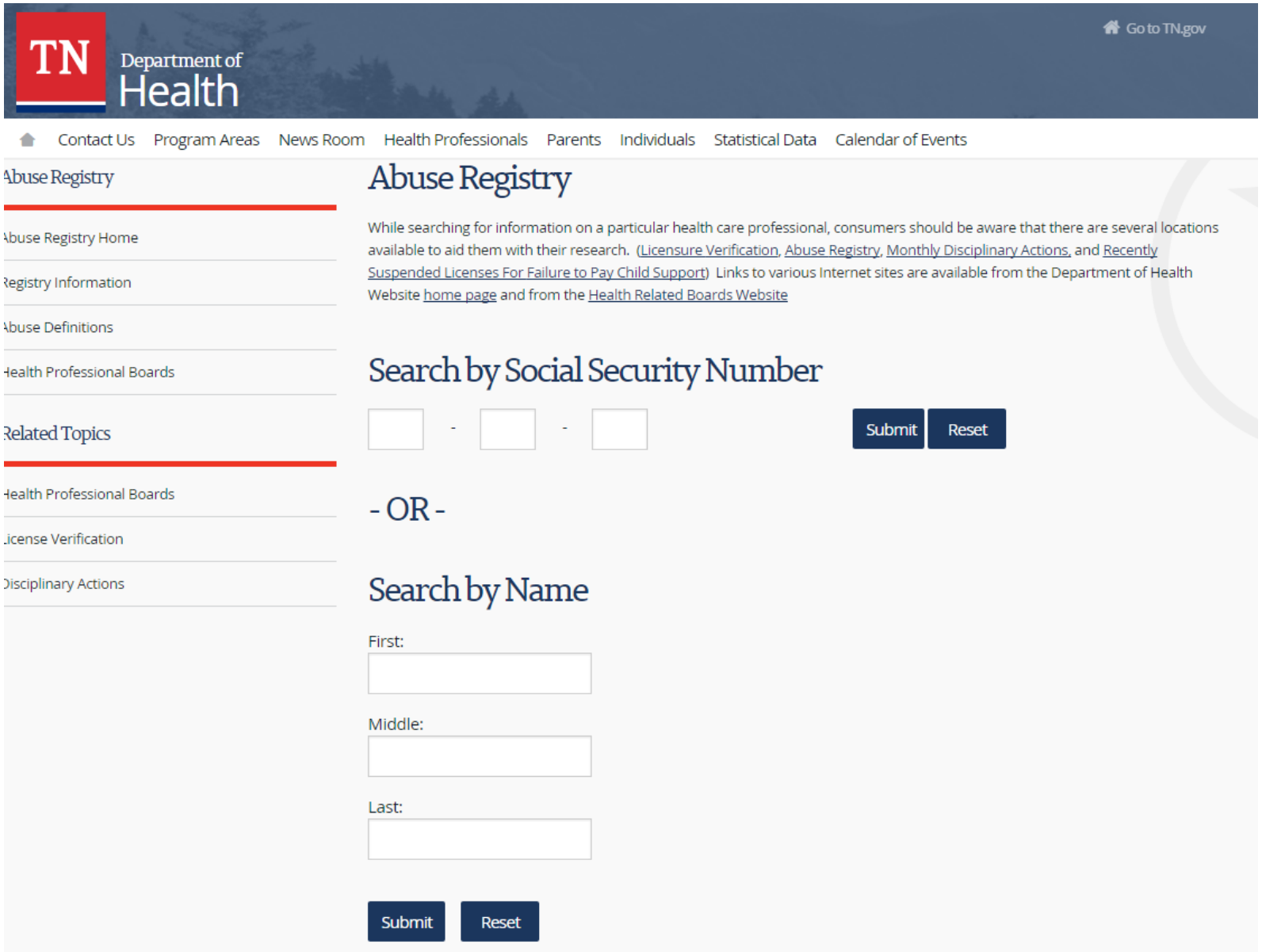
Serial Numbers: ZZZZ

- Last 4 digits
- SNs have been issued “in monotonically increasing order” within each State and within each GN
  - From 0001 to 9999
- However, SSA also writes:

  *“SSNs are assigned **randomly** by computer within the confines of the area numbers allocated to a particular state based on data keyed to the Modernized Enumeration System” (50, RM00201.060)*

- …reflecting SSA’s belief that idiosyncrasies in the SSN application and vetting processes make the SN assignment effectively random

Abuse Registry



https://apps.health.tn.gov/abuseregistry/

HSBC The world's local bank

1 Provide Your Information — 2 Agree to Terms — 3 View Your Status — 4 You're Done!

Helpful Hint: You can use the TAB key to move between fields

Terms & Conditions

Personal Information

First Name
MI
Last Name
Suffix None
Social Security Number - -
Date of Birth(MM/DD/YYYY) - -
Email Address
Re-Enter Email Address
Mother's Maiden Name(Last Name Only. example; Smith or Jones)

Address Information

Current Address
Apt.#,Bldg.#, etc
City
State
Zip Code

Address Information

Current Address
Apt.#,Bldg.#, etc
City
State
Zip Code
Residence Type
Monthly Mortgage/Rent Amount
Home or Cell Phone Number - -

Employment Information

Employer Name
Work Phone Number - -
Annual Income*

* Alimony, child support and separate maintenance income need not be disclosed if you do not wish to have it considered as a basis for repaying this obligation.

Online Validation

- http://www.ssnvalidator.com

Chances of correctly matching SSN digits by random guess, under *status quo* knowledge

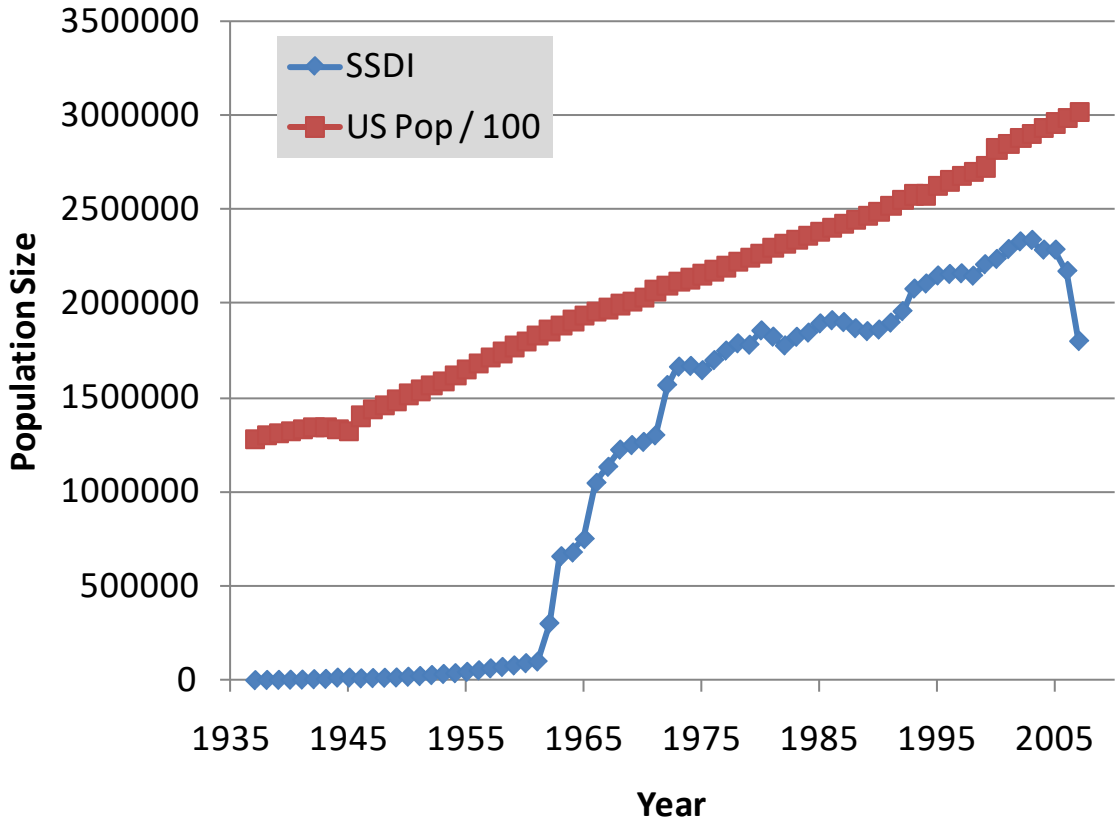
	Alaska, 1998		New York, 1998	
	First 5 digits with 1 guess	All 9 digits with < 1,000 guesses	First 5 digits with 1 guess	All 9 digits with < 1,000 guesses
No auxiliary knowledge	0.0014%	0.00014%	0.0014%	0.00014%
Knowledge of state of SSN application	1%	0.1%	0.012%	0.0012%

Or SSDI

- You could wait until someone dies…
- Social Security Death Index (SSDI) Database
  - http://search.ancestry.com/search/db.aspx?dbid=3693
- Death reported to the Social Security Administration
  - Possibly submitted by “relation” requesting Social Security benefits
    http://www.ssa.gov/pubs/10084.html
  - or to stop benefits

Rise of the SSN

- Over 80 million records
- Data back to 1937, but the majority is after 1962



Reasons to believe that the assignment lacks sufficient randomness

- In the last 30 years, SSN issuance has become more regular
  - Increasing computerization of the public administration, including SSA and its various fields offices
  - After 1972, SSN assignment centralized from Baltimore, MD
  - After 1989, Enumeration at Birth Process (EAB)
    - Prior to 1989, only small percentage of people received SSN when they were born
    - Currently at least 90 percent of all newborns receive SSN via EAB together with birth certificate

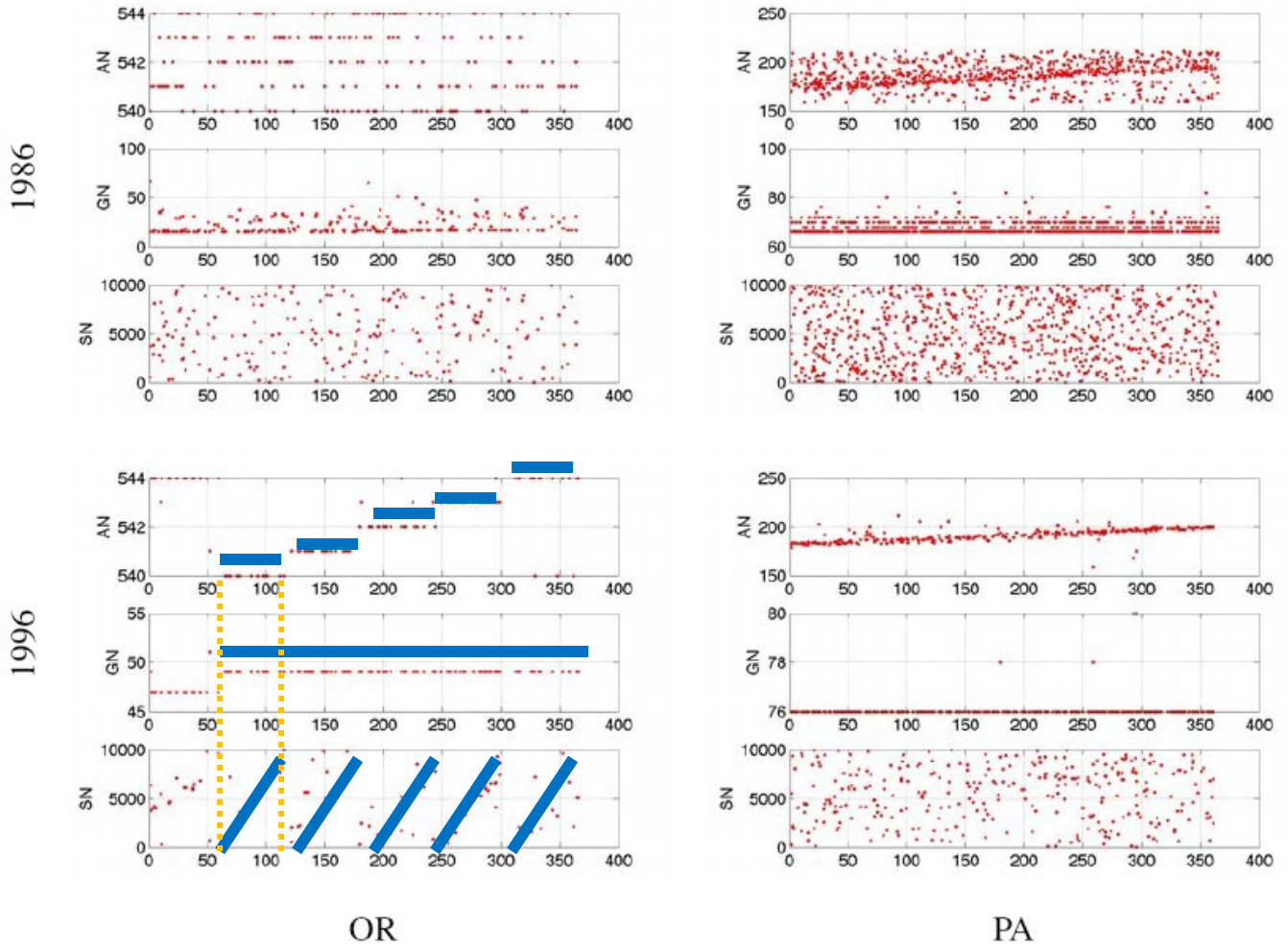
Adapted from A. Acquisti

Hence, two hypotheses

1. Expect SSN issuance patterns to have become more regular over the years, i.e., increasingly correlated with an individual's birthday and birthplace
   - This should be detected through analysis of available SSN data

2. Expect these patterns to have become so regular that it is possible to infer *unknown SSNs* based on the patterns detected on *available SSNs*
   - This should be verified by contrasting estimated SSNs against known SSNs

SSN assignment patterns: Two representative States



Prediction Approach

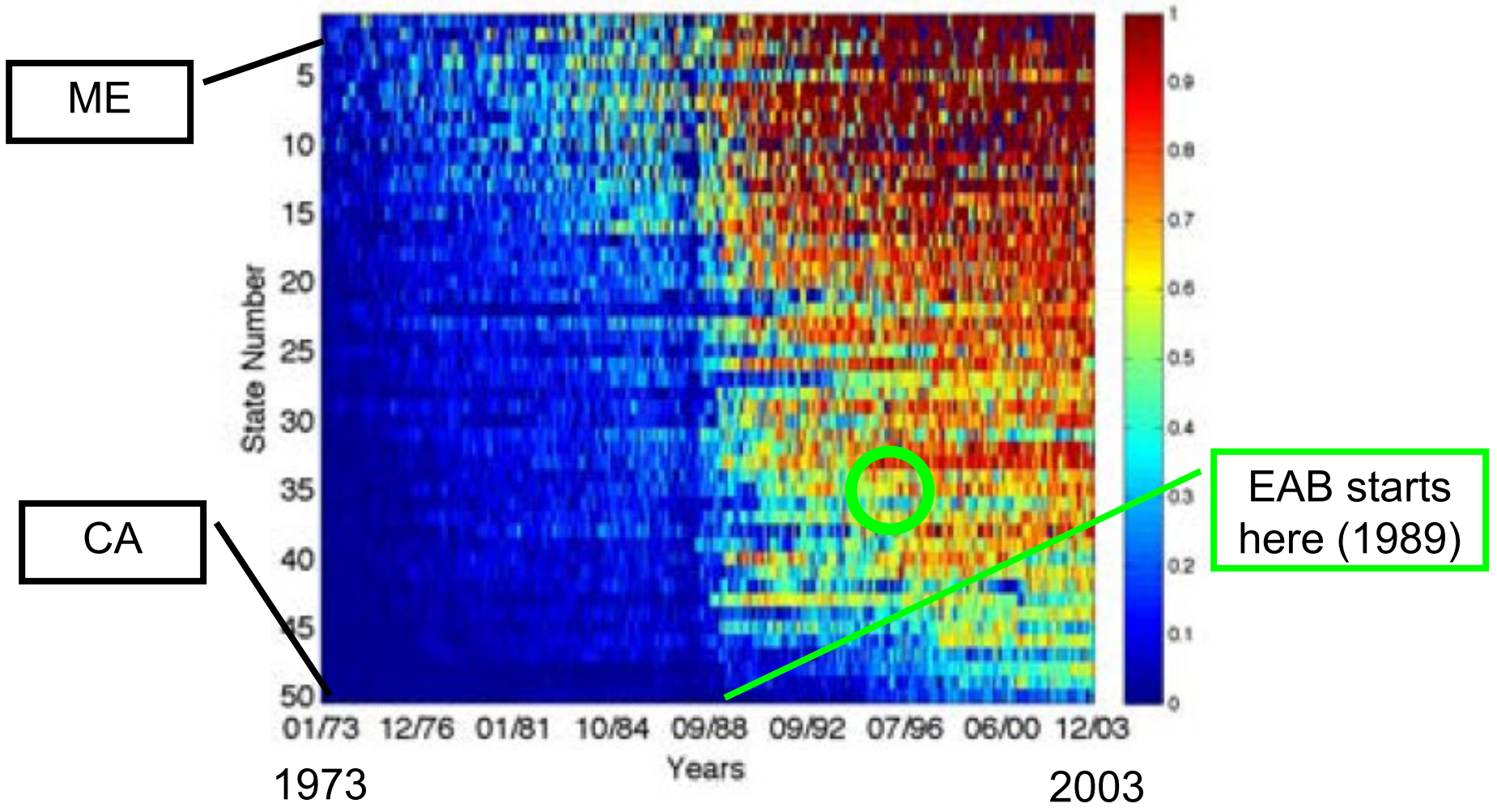
- Area number
  - Mode AN in target's state around target's birthday
- Group number
  - Mode GN in target's state around target's birthday
- Serial number

  - Based on regressions coefficients, inserting target's birthday as *dd*

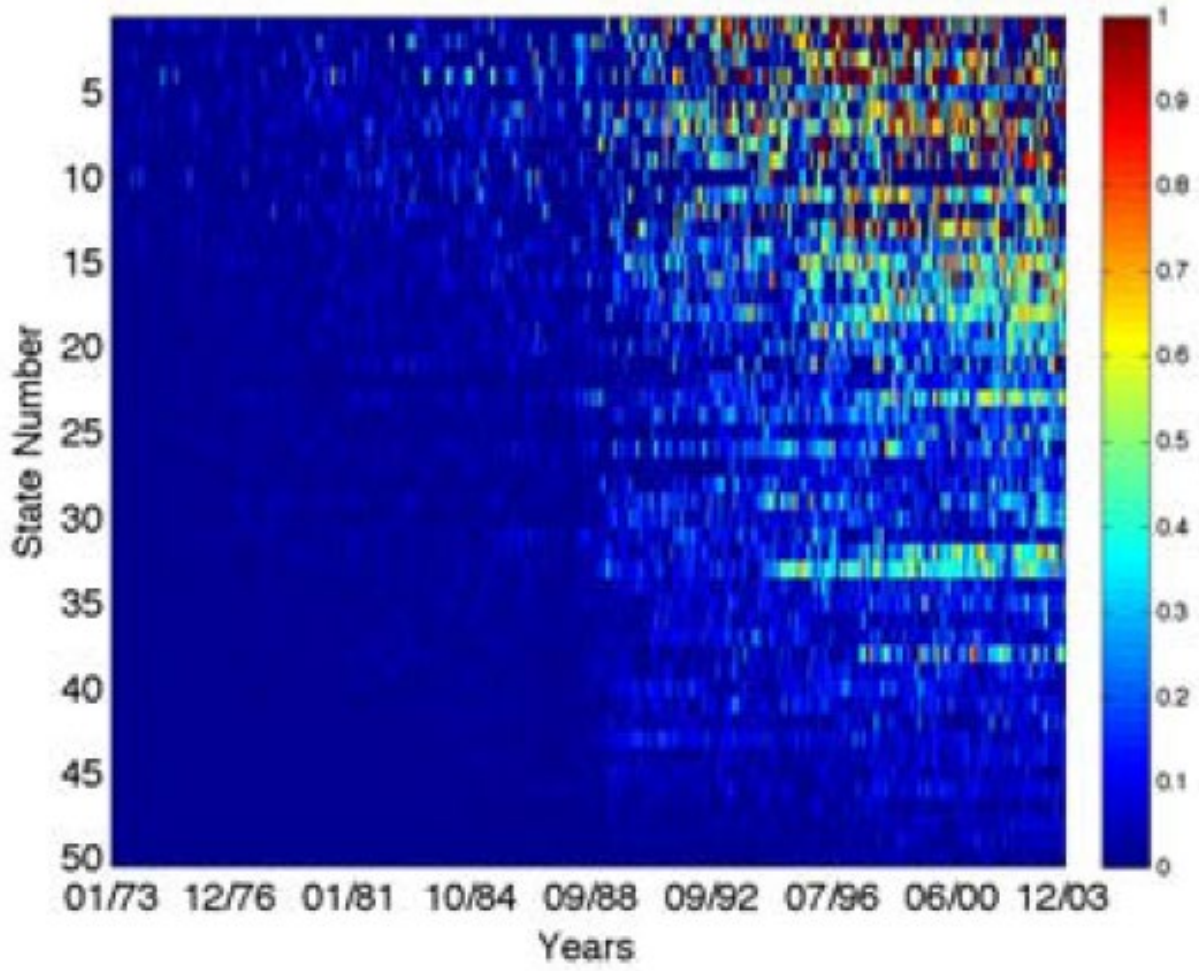
Success metrics

- Accuracy in prediction of the first 5 digits of an individual's SSN with 1, 10, 100, and 1000 attempts
  - Note: 1,000 attempts is equivalent to 3-digit PIN
  - And is very insecure and vulnerable to brute force attacks

AN-GN predictability (first 5 digits)



Full SSN predictability with <1,000 attempts



Test 2: From social networks data to SSNs

- Used birthday data of 621 alive individuals to predict their SSN, based on interpolation with DMF data
- Sample: born in 1986-1990 (i.e., mostly before EAB)
- In most populous states (i.e., worst case scenario)

- Birthday and birthplace data can be obtained from several sources, but most easily, and in mass amounts, from online social networks

Facebook estimations

- Overall AN prediction accuracy: 8.5%
- Overall GN prediction accuracy: 29.1%
- Combined AN-GN prediction accuracy: 6.3%
  - Compare to corresponding weighted sample in Test 1 (based on DMF data): 11.21%

Clinical Text Privacy Risks

Wednesday, February 2, 1994

Marjorie Long, M.D.
St. John's Hospital
Huntington 18
Boston, MA 02151

RE: Virginia Townsend
CH#32-841-09787
DOB 05/26/86

Dear Dr. Lang:

I feel much better after seeing Virginia this time. Dot is a 7 and 6/12 year old female in follow up for insulin dependent diabetes mellitus diagnosed in June of 1993 by Dr. Frank at Brigham's. She is currently on Lily Human Insulin and is growing and gaining weight normally. She will start competing again with the U. S. Junior Gymnastics team. We will contact Mrs. Hodgkins at Marina Corp 473-1214 for a follow-up visit for her daughter.

Patrick Hayes, M.D. 34764

“Scrubbed” Document

Erisa Cosborn
Brighaul Hos
Alberdam W
Peabon, MA

Dear Dr. ***Jandel***:

I feel much better after seeing ***Kathel*** this time. ***Cob*** is a 7
year old fema
mellitus diag
She is curren
gaining weigh
___ . We will er daughter.

Does this satisfy Safe Harbor?

Does this satisfy a Limited Dataset?

Scrub (Sweeney '96)

- Itemizes Commonsense knowledge
  - What does first/last name, phone #, … look like?
  - Provide lists of names, nicknames, abbreviations and parser/algorithms extract templates
- Detection based on precedence and probability via template observation / combination
- Identifier Detection based on AI model
  - Blackboard architecture integrates detectors to systematically discover identifiers

Sweeney L. Replacing Personally-Identifying Information in Medical Records, the Scrub System. In Proceedings of the 1996 AMIA Annual Symposium. 1996: 333-337.

Facial Image Privacy Risks

- Bioterrorism Surveillance Problem

"traditional" for bioterrorism surveillance

extended list of sources

very non-traditional data

Data sources
hospital inpatient claims
emergency room visits
physician office visits
lab reports
prescriptions
assisted living reports
school absentism
business absentism
animal reports
grocery purchases
web clickstream data
death data
videos of public spaces
GPS auto data

Tracking Faces in Video Data

- The world isn't a static place
- Video and automated tracking of people is already here

  … and getting better.

Camera-Happy World

- Over 85 million cameras in the US
  - Manhattan has over 15,000 cameras
  - Average American caught on camera 70 times / day

- Over 500,000 cameras in London
  - Over 15,000 in the underground alone

How to de-identify facial images?

Readings for this Friday

- 1. Sweeney L. **k-anonymity: A model for protecting privacy**. *International journal of uncertainty, fuzziness and knowledge-based systems*. 2002 Oct;10(05):557-70.

- Optional
  - ❑ 2. Dankar FK, El Emam K. **A method for evaluating marketer re-identification risk**. *InProceedings of the 2010 EDBT/ICDT Workshops* 2010 Mar 22 (pp. 1-10).
  - ❑ 3. Newton EM, Sweeney L, Malin B. **Preserving privacy by de-identifying face images**. *IEEE transactions on Knowledge and Data Engineering*. 2005 Jan 10;17(2):232-43.

Feedback Survey

- One thing you learned or felt was valuable from today's class & reading

- Muddiest point: what, if anything, feels unclear, confusing or "muddy"

- https://www.wjx.cn/vm/hX0mIro.aspx