ACDEB

Advisory Committee on Data for Evidence Building

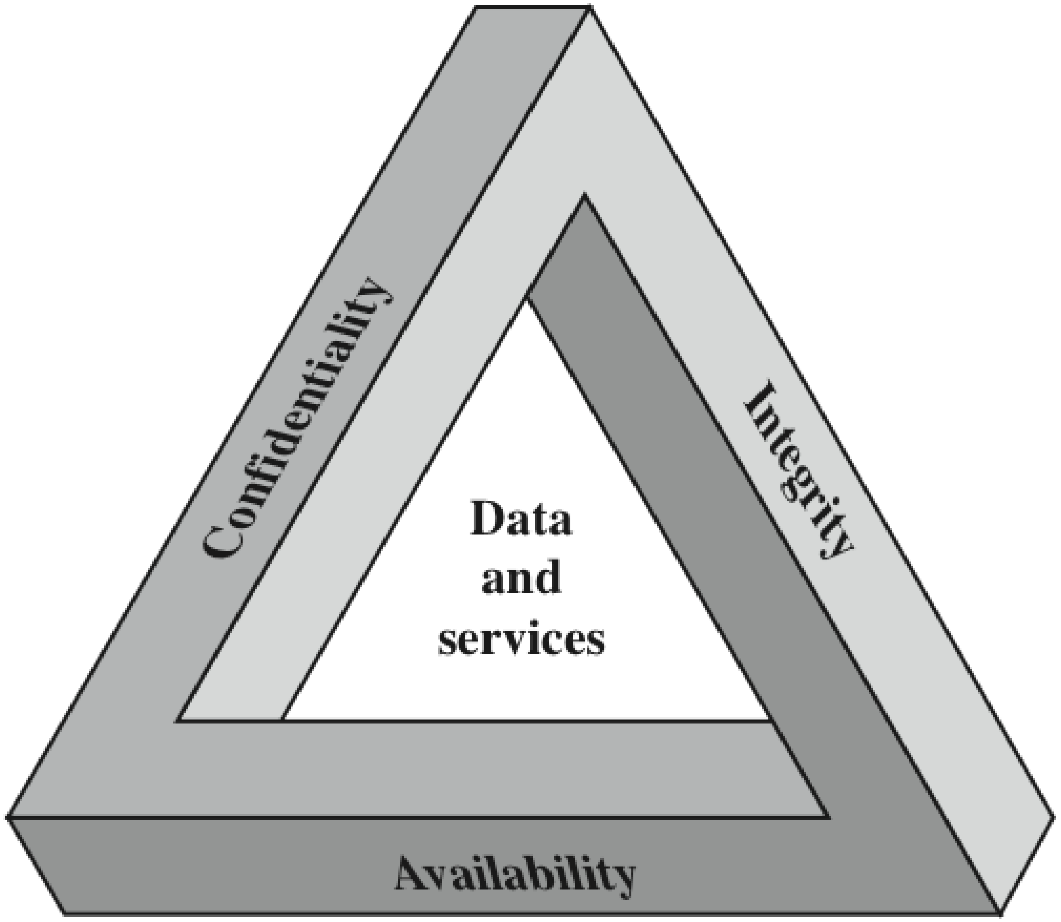
Privacy and Confidentiality: Concepts & Challenges

Shawn Davis

Director of Digital Forensics – Edelson PC

Adjunct Industry Professor – Illinois Tech

Right of an individual to have control over how his or her personal information is collected, used, and/or disclosed



Core Security Principles

Privacy Act of 1974

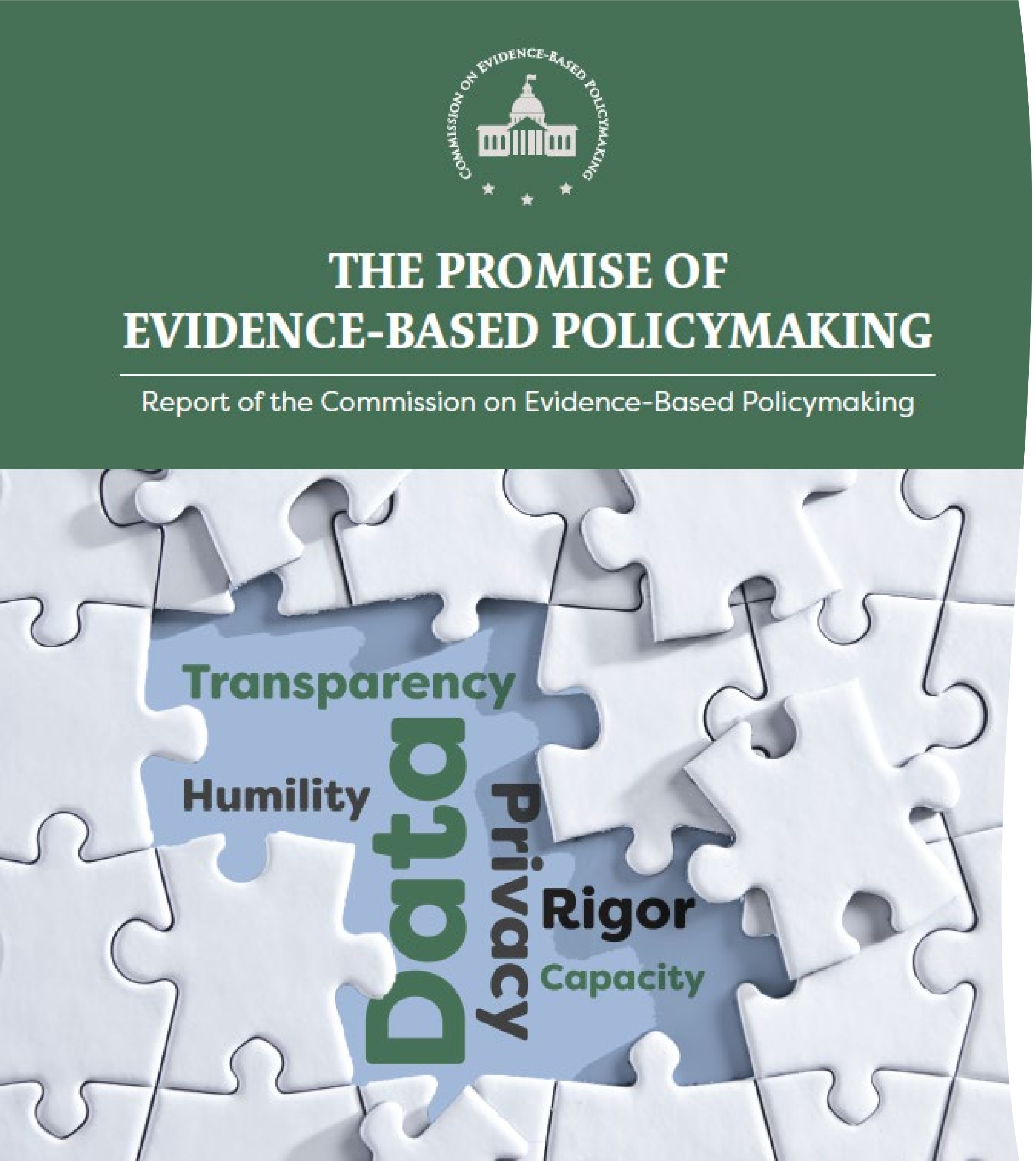
No agency shall disclose any record which is contained in a system of records by any means of communication to any person, or to another agency, except pursuant to a written request by, or with the prior written consent of, the individual to whom the record pertains, unless disclosure of the record would be—

(5) to a recipient who has provided the agency with advance adequate written assurance that the record will be used solely as a statistical research or reporting record, and the record is to be transferred in a form that is **not individually identifiable**;

Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA)

Data or information acquired by an agency under a pledge of confidentiality for exclusively statistical purposes shall not be disclosed by an agency in **identifiable form**, for any use other than an exclusively statistical purpose, except with the informed consent of the respondent

(4) The term **"identifiable form"** means any representation of information that permits the **identity of the respondent** to whom the information applies to be **reasonably inferred by either direct or indirect means**.

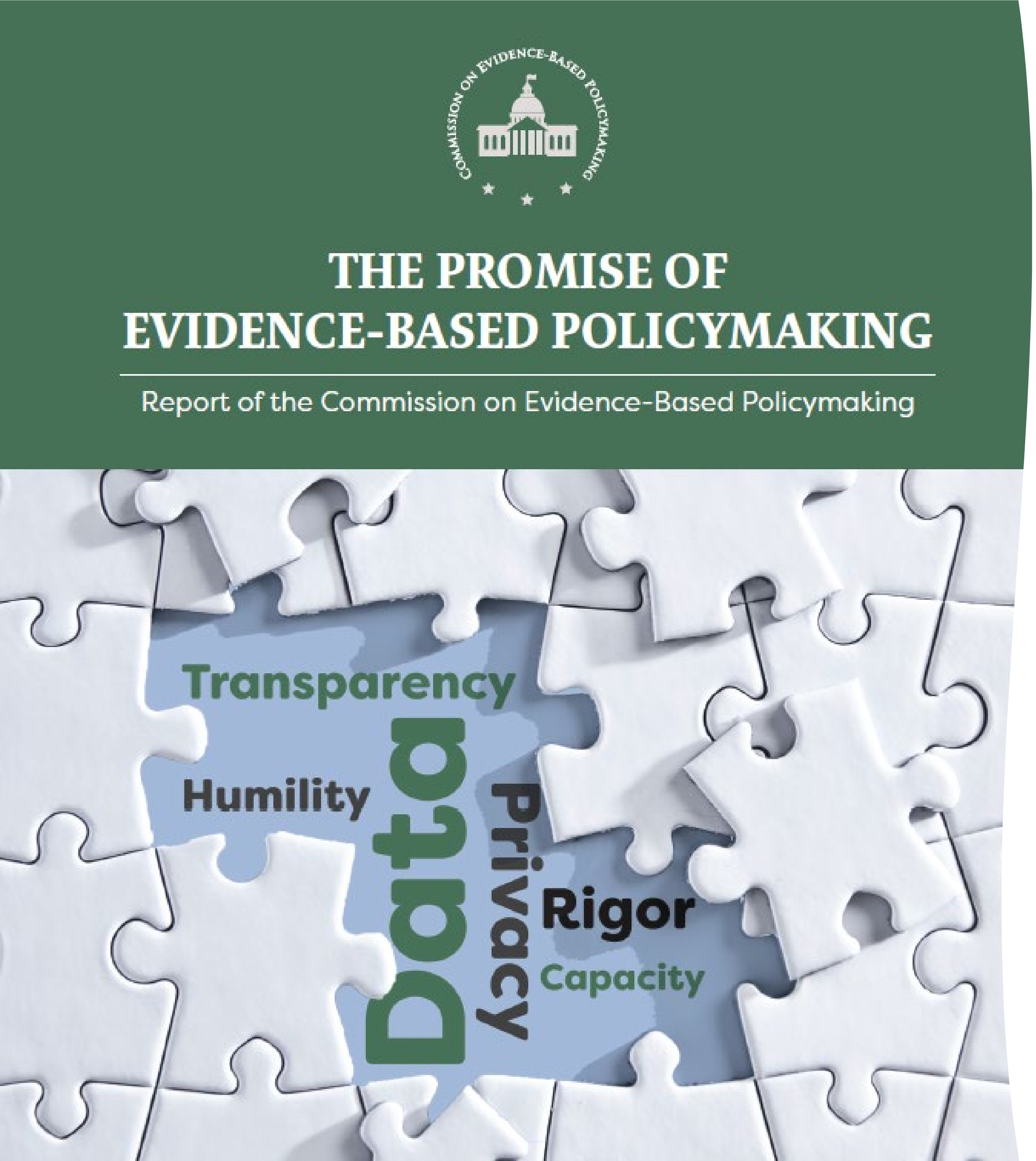


THE PROMISE OF EVIDENCE-BASED POLICYMAKING

Report of the Commission on Evidence-Based Policymaking

Recommendations

1. Promote public trust in accuracy of information used to guide government decision-making
2. Establish a new transparency and accountability portal to ensure the public is notified about how confidential data is used
3. Develop a uniform process for external researchers to apply and qualify for secure access



COMMISSION ON EVIDENCE-BASED POLICYMAKING

THE PROMISE OF EVIDENCE-BASED POLICYMAKING

Report of the Commission on Evidence-Based Policymaking

Main Privacy Recommendations

1. Amend Privacy Act and CIPSEA to require departments to conduct risk assessments for public releases of **de-identified confidential data**

2. Providing secure and restricted access to confidential data

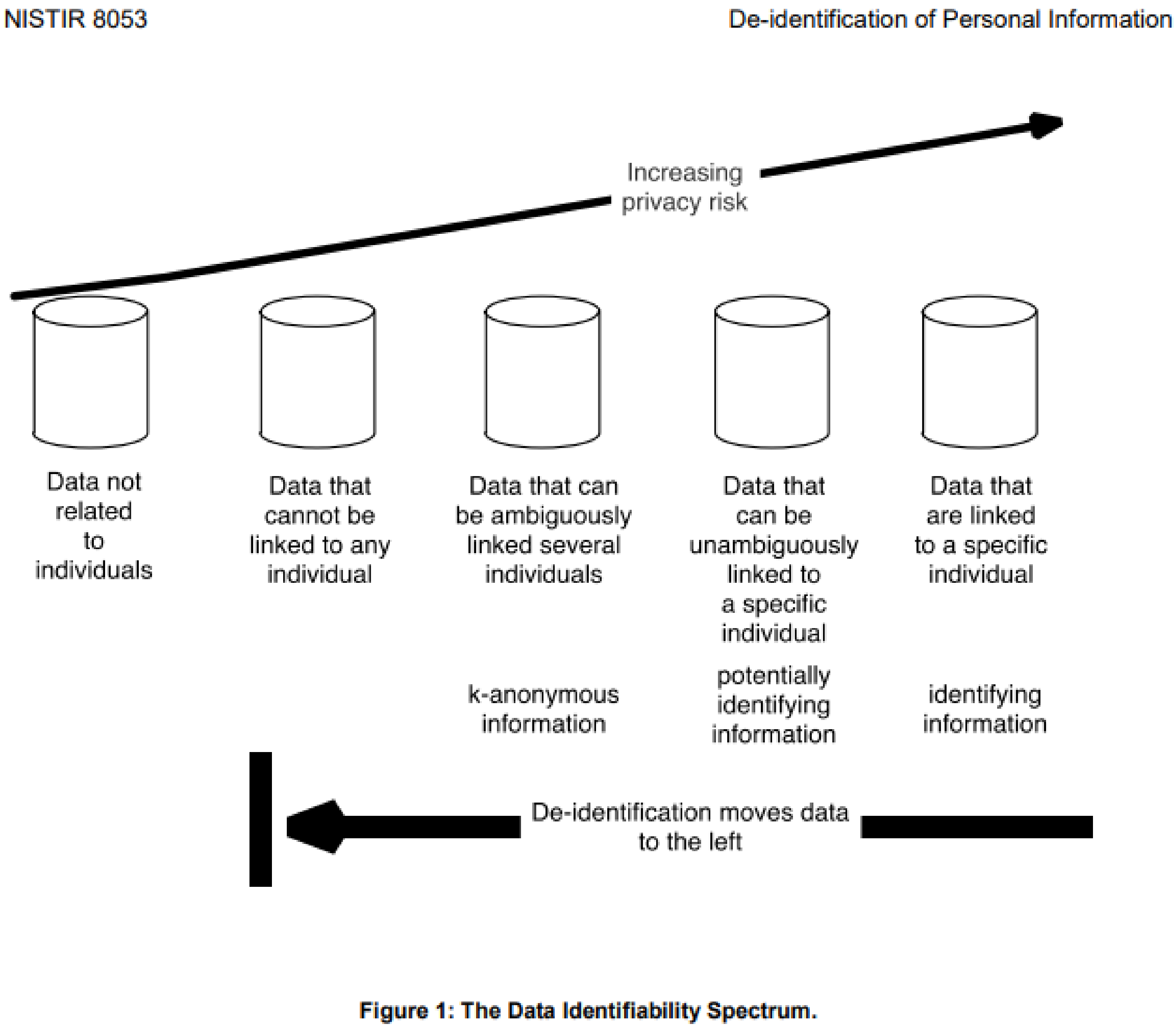
3. Adoption of cutting-edge technology for data security, integrity, and confidentiality.

Personally Identifiable Information (PII)

- **Direct Identifiers**
  - Name
  - Address
  - SSN
  - Phone Numbers
  - Biometrics
  - Email address
  - Account numbers
  - License numbers
  - Vehicle identifiers
  - Device identifiers
  - IP addresses
  - Photographs
  - URLs
  - Etc.

- **Indirect Identifiers**
  - Location
  - Ethnicity
  - Race
  - Religion
  - Age
  - Zip Code
  - DOB
  - Gender
  - Financial transactions
  - Place of birth
  - Medical information
  - Etc.

NISTIR 8053 De-Identification of Personal Information



NISTIR 8053

De-identification of Personal Information

Increasing privacy risk

Data not related to individuals

Data that cannot be linked to any individual

Data that can be ambiguously linked several individuals

Data that can be unambiguously linked to a specific individual

Data that are linked to a specific individual

k-anonymous information

potentially identifying information

identifying information

De-identification moves data to the left

Figure 1: The Data Identifiability Spectrum.

https://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf

“De-identified” Terminology Confusion

Dept of Education: “…removal of all personally identifiable information”

HIPAA: “…no reasonable basis to believe that the information can be used to identify an individual”

CCPA: “…cannot reasonably identify, relate to, describe, be capable of being associated with, or linked, directly or indirectly, to a particular customer”

Data Category Examples

Personally Identifiable Data (Absolute or High Risk)

Record ID	Name	SSN	DOB	Gender	Income	Job	Height	Zip	Geolocation
12548	Kat Robin	123-45-6789	01/02/60	F	60,000	Teacher	6’0”	60302	41.8869,-87.7801
12567	Jim Jones	987-65-4321	05/02/72	M	65,000	Teacher	5’11”	60646	41.9945,-87.5845
	Direct	*Direct*	*Indirect*	*Indirect*				*Indirect*	*Indirect*

Pseudonymized Data (Medium Risk)

Record ID	Name	SSN	DOB	Gender	Income	Job	Height	Zip	Geolocation
12548	sadfjkwESdi	29j2k34kjw	01/02/60	M	60,000	Teacher	6’0”	60302	41.8869,-87.7801
12567	Ekriudlkjwe	09dfgk3453	05/02/72	M	65,000	Teacher	5’11”	60646	41.9945,-87.5845
	Pseudonym	Pseudonym	*Indirect*	*Indirect*				*Indirect*	*Indirect*

De-identified Data (Low Risk)

Record ID	Income	Job	Height
12548	60,000	Teacher	6’0”
12567	65,000	Teacher	5’11”

Anonymized Data [Summary example] (Zero Risk)

Count	Income	Job	Height
2	60-65k	Teacher	5’11” to 6’0”

Re-identification Risks – Combination of Indirect Identifiers

Record ID	Name	SSN	DOB	Gender	Income	Job	Height	Zip	Geolocation
12548	sadfjkwESdi	5484548	01/02/60	M	60,000	Teacher	6’0”	60302	41.8869,-87.7801
	Pseudonym	Pseudonym	*Indirect*	*Indirect*				*Indirect*	*Indirect*

“It was found that 87% (216 million of 248 million) of the population in the United States had reported characteristics that likely made them unique based only on {5-digit ZIP, gender, date of birth}.”

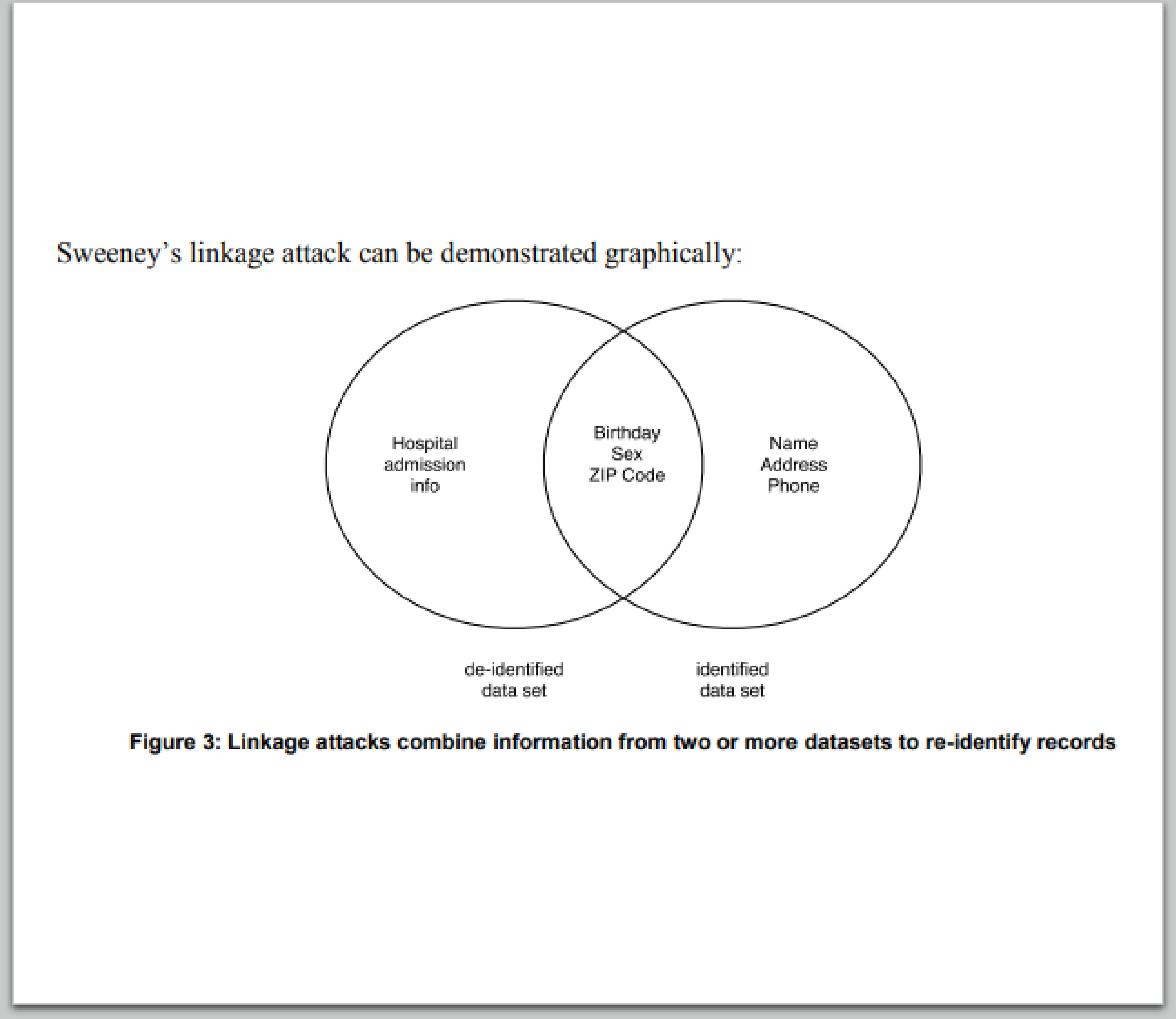
https://dataprivacylab.org/projects/identifiability/paper1.pdf

Re-identification Risks: Linkage Attack Using Indirect Identifiers

MIT Grad Student Latanya Sweeney re-identified MA Gov. Weld from two datasets:

1. Insurance reimbursement records

2. Voter registration list



Sweeney's linkage attack can be demonstrated graphically:

Figure 3: Linkage attacks combine information from two or more datasets to re-identify records

https://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf

Other Real World Linkage Attack Examples

Netflix Prize: Researchers matched de-identified Netflix viewing data with IMDB data.

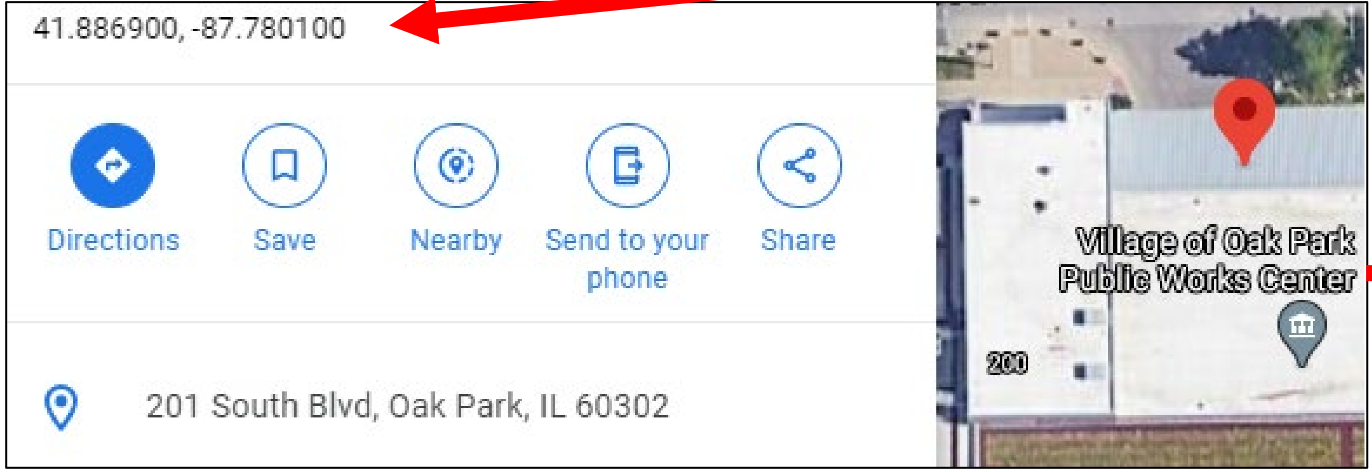
Medical Tests: 5-7 lab results enough to identify a patient from de-identified biomedical research database.

Credit Card Transactions: Uniquely identified 90% of people in de-identified credit card transactions from four distinct transactions in space and time.

Taxi Ride Data: NYC de-identified dataset of 173 million taxi rides and ride timestamps, but left 32-bit code (that could be easily converted to a taxi medallion number). Individuals re-identified from photos of themselves and taxi medallion along with photo timestamp.

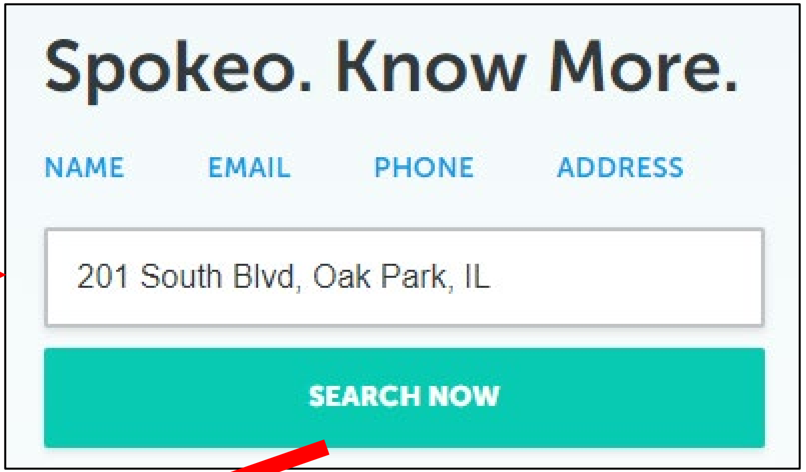
Re-identification Risks - Geolocation

Record ID	Name	SSN	DOB	Gender	Income	Job	Height	Zip	Geolocation
12548	sadfjkwESdi	5484548	01/02/60	M	60,000	Teacher	6’0”	60302	**41.8869,-87.7801**
	Pseudonym	Pseudonym	*Indirect*	*Indirect*				*Indirect*	***Indirect***



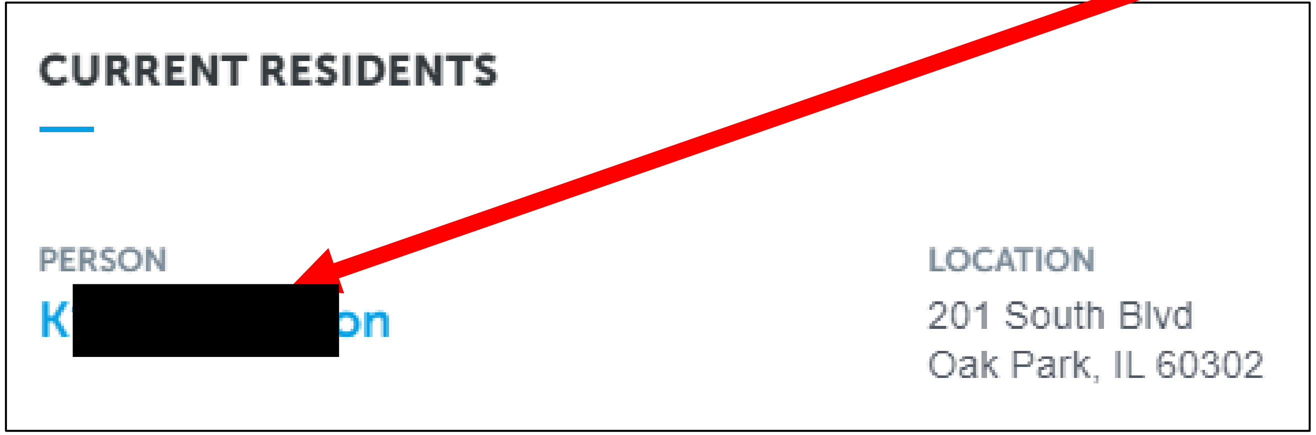
41.886900, -87.780100

201 South Blvd, Oak Park, IL 60302



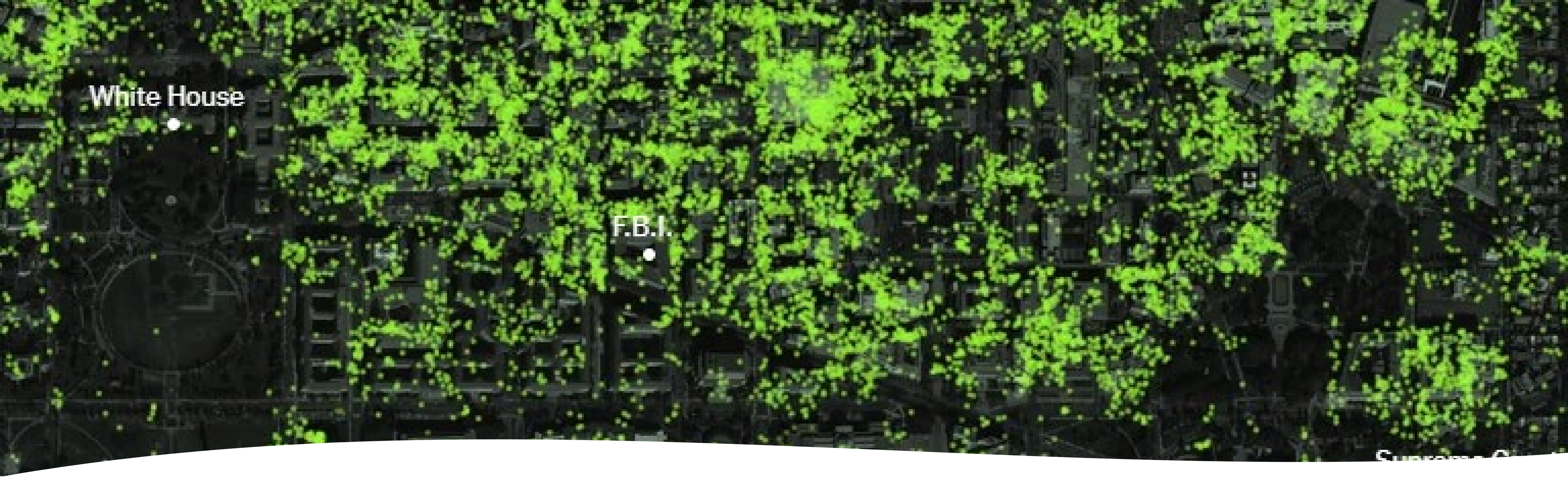
Spokeo. Know More.

201 South Blvd, Oak Park, IL



CURRENT RESIDENTS

PERSON	LOCATION
K[redacted]on	201 South Blvd Oak Park, IL 60302

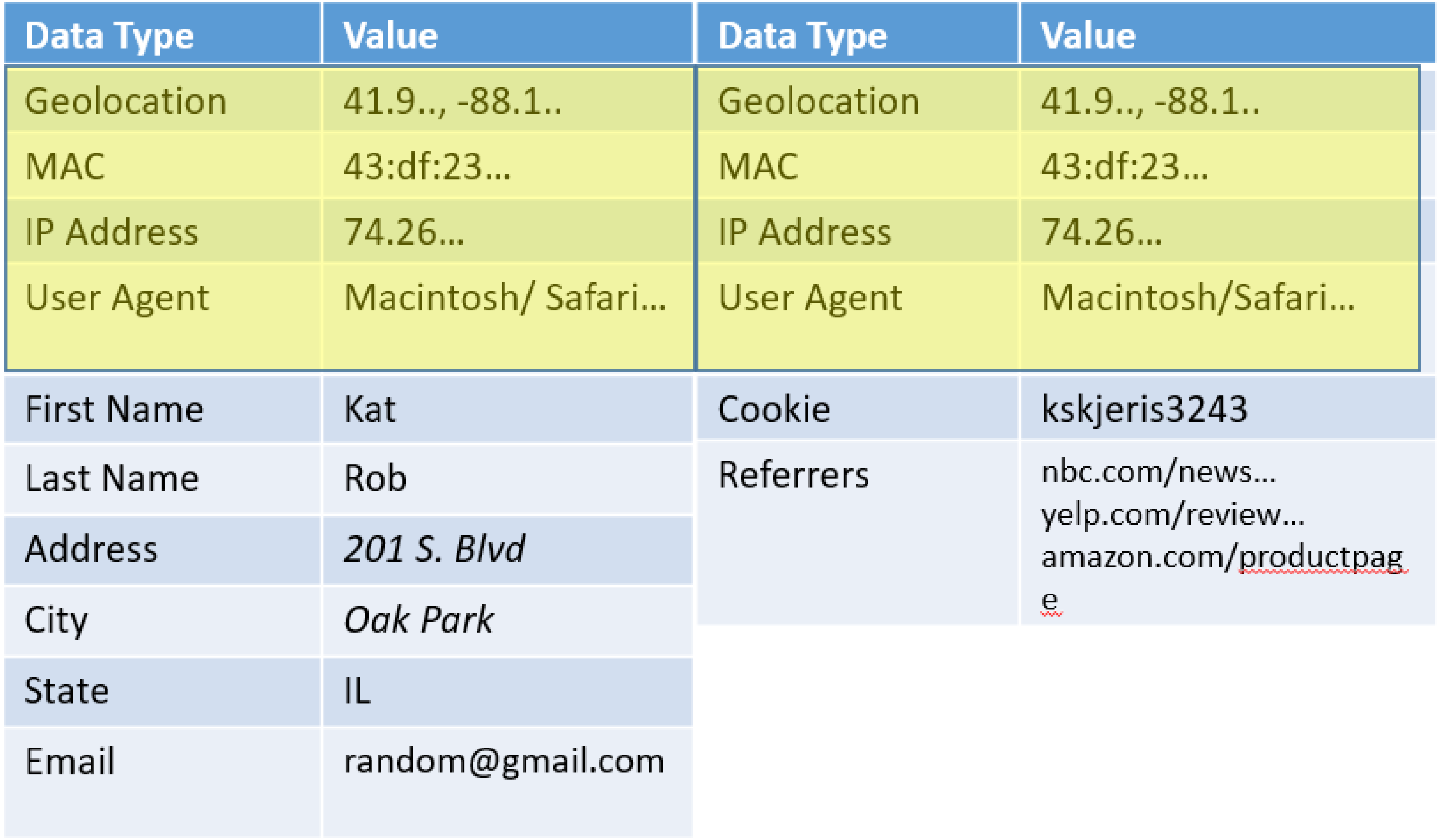


NYT 2019 Review of Geolocation DB – 12 Mil Phones

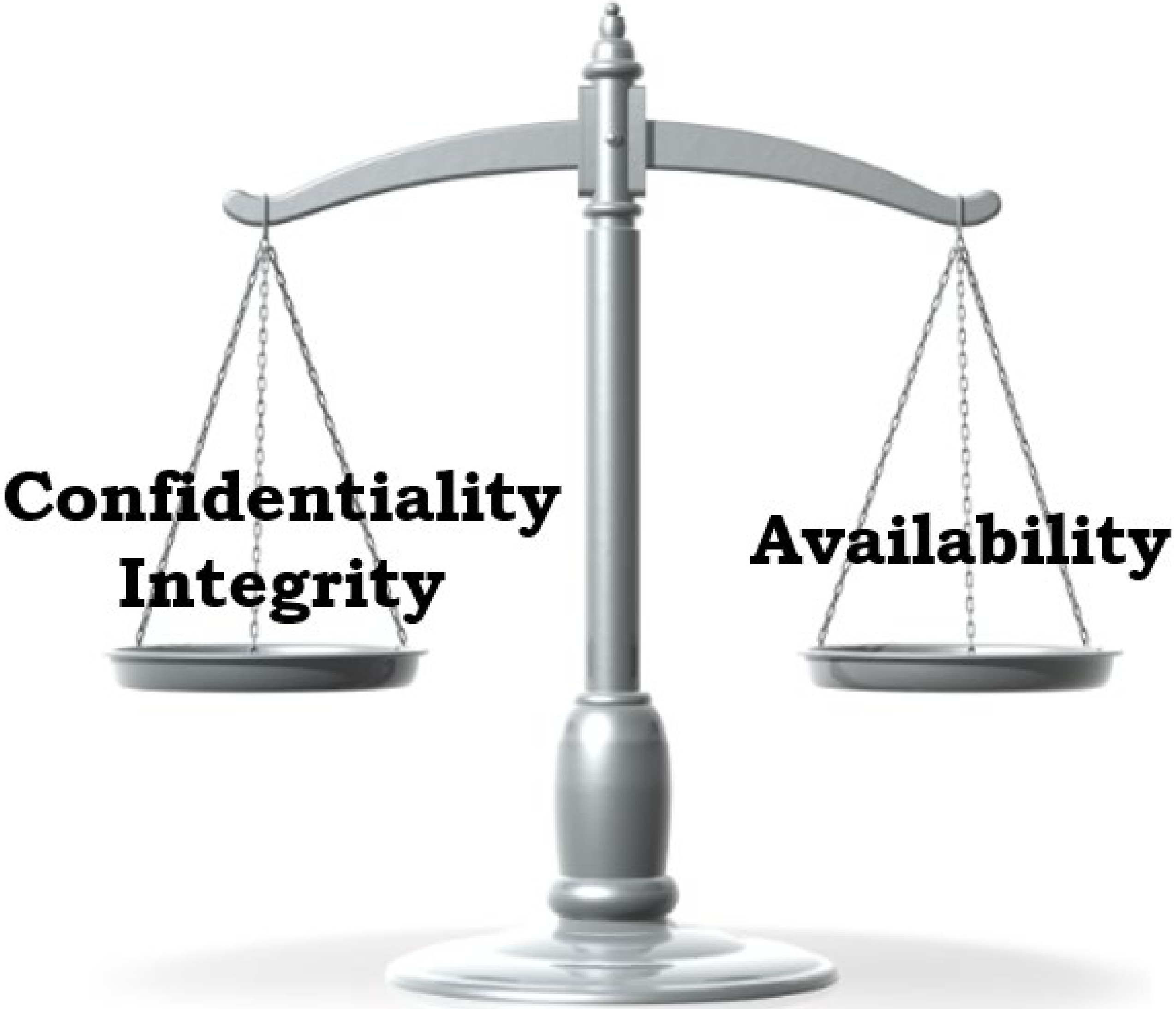
"We were able to track smartphones in nearly every major government building and facility in Washington."

"We could follow them back to homes and, ultimately, their owners' true identities."

Re-identification Risks – Matching Device Metadata



Data Type	Value	Data Type	Value
Geolocation	41.9.., -88.1..	Geolocation	41.9.., -88.1..
MAC	43:df:23…	MAC	43:df:23…
IP Address	74.26…	IP Address	74.26…
User Agent	Macintosh/ Safari…	User Agent	Macintosh/Safari…
First Name	Kat	Cookie	kskjeris3243
Last Name	Rob	Referrers	nbc.com/news… yelp.com/review… amazon.com/productpage
Address	*201 S. Blvd*		
City	*Oak Park*		
State	IL		
Email	random@gmail.com		


Confidentiality
Integrity
Availability