Data Science at the Intersection of Privacy and Security

UC Berkeley, Master in Data Science and Information
W231 Spring 2016
Alfred Arsenault & Tarun Chopra

An approach to balancing **security** and **privacy** through **Data Science.**

Introduction and Context

1

Security and **Privacy**, similar but different.

2

Finding right balance between Security and Privacy is critical for **Organizations**.

3

Data Science can help or make it harder to achieve.

US and State Governments

Security
(Four Pillars)

Confidentiality

Integrity

Availability

Accountability

Privacy
(Facets governing privacy)

Anonymity - Concealing the identity of an entity.

Pseudonymity - Concealing the true identity of an entity. E.g, “Publius”, “Satoshi Nakamoto”

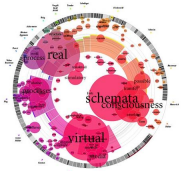
Fairuse- Data collection relevant to the context.

Access- Limited access to data.

Control Aggregation- Restricted distribution and appropriate sanitization.

Control Lifecycle- Data operationalization management.

Data Science (Multilayered Discipline)



Examining raw data and drawing **meaningful conclusions**.

Analyze historical and current facts to make **predictions about future** and unseen situations.

Grasp difficult concepts or **identify new patterns**.

Systems "learn" from previous data; generalize to unseen samples.

Draw conclusions about people and systems from analysis of **observed data**.

Organization decision around Security and Privacy

E.g.: Choose Accountability or Anonymity

- Defense Department bans anonymity on official systems

US Government: NIST Special Publication 800-53 lists all security controls, including privacy controls

Formal guidance on how to choose security controls - FIPS 199 & Chapter 3 of 800-53

Privacy controls are less formal - provided to Chief Privacy Officers/Senior Agency Officials for Privacy

Work with program managers on a case-by-case basis to determine privacy requirements

Security and Privacy Controls for Federal Information Systems and Organizations, NIST Special Publication 800-53 Revision 4, National Institute of Standards and Technology, April 2013

Privacy and Security - Synergy and Interference

Privacy -> Security	Anonymity	Pseudonymity	Controlled lifecycle	Fair Use	Access	Control of Aggregation
Confidentiality	Support	Support	Support	Support	Support	Harm
Integrity	Neutral	Neutral	Neutral	Support	Support	Harm
Availability	Neutral	Neutral	Harm	Harm	Harm	Support
Accountability	Prevent	Harm	Support	Support	Support	Harm

What Does Data Science Have to Do With It?

Impact of Data Science on Privacy

	Anonymity	Pseudonymity	Controlled lifecycle	Fair Use	Access	Control of Aggregation
Statistical Analysis	Harm	Harm	Neutral	Neutral	Neutral	Harm
Machine Learning	Neutral	Neutral	Neutral	Neutral	Neutral	Harm
Data Analytics	Harm	Harm	Neutral	Neutral	Neutral	Harm
Predictive Analytics	Harm	Harm	Neutral	Neutral	Neutral	Harm
Data Visualization	Harm	Harm	Neutral	Neutral	Neutral	Harm

Note: if the Data Science technique both* Supports *and* Harms *the Privacy aspect the overall rating is "Neutral"

Impact of Data Science on Security

	Confidentiality	Integrity	Availability	Accountability
Statistical Analysis	Harm	Neutral	Neutral	Support
Machine Learning	Neutral	Neutral	Neutral	Neutral
Data Analytics	Harm	Neutral	Neutral	Support
Predictive Analytics	Harm	Neutral	Neutral	Support
Data Visualization	Neutral	Neutral	Neutral	Neutral

Note: if the Data Science technique both* Supports *and* Harms *the Security aspect the overall rating is "Neutral"

Examples

Example 1: Data Aggregation Harms Anonymity and Confidentiality

- Denver crime database shows home address where crime occurred
- Cross-reference against property tax database and voter registration database to determine identity of residents
    - The voter registration database is public information but costs $1,000

Date of Crime	Type of Crime	Address
2/7/1997	Domestic Assault	3301 N. Krameria

3301 KRAMERIA ST

Owner	Schedule Number	Legal Description	Property Type	Tax District
WORTHAM,ASFORD Q 3301 KRAMERIA ST DENVER , CO 80207-2137	0129220030000	OAKLAND B21 L25 & 26	RESIDENTIAL DUPLEX	DENV

Summary Assessment Assessment Protest Taxes Comparables Neighborhood Sales Chain of Title

Example 1 (cont.)

3301 KRAMERIA ST

Owner	Schedule Number	Legal Description	Property Type	Tax District
WORTHAM,ASFORD Q 3301 KRAMERIA ST DENVER , CO 80207-2137	0129220030000	OAKLAND B21 L25 & 26	RESIDENTIAL DUPLEX	DENV

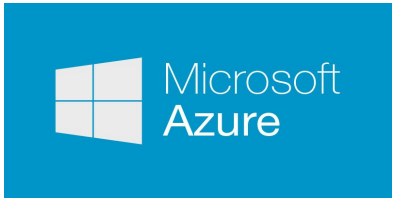
Summary Assessment Assessment Protest Taxes Comparables Neighborhood Sales Chain of Title

Chain Of Title Records

Reception Number	Reception Date	Instrument	Sale Date	Sale Price	Grantor/ Grantee
2015029413	3/9/2015	QC	3/6/2015	$10	APPLETON,WANDA/ WORTHAM,ASFORD Q
0000229434	12/4/2002	DC	9/24/1995		BUSBY,FANNIE &/ APPLETON,WANDA
0000109529	7/8/1994	WD	6/30/1994	$89,000	GAYLES,FREDDIE &/ BUSBY,FANNIE &
0000023888	9/22/1986	WD	9/22/1986		BROWN,LEROY F & CHARLOTTE/ GAYLES,FREDDIE &

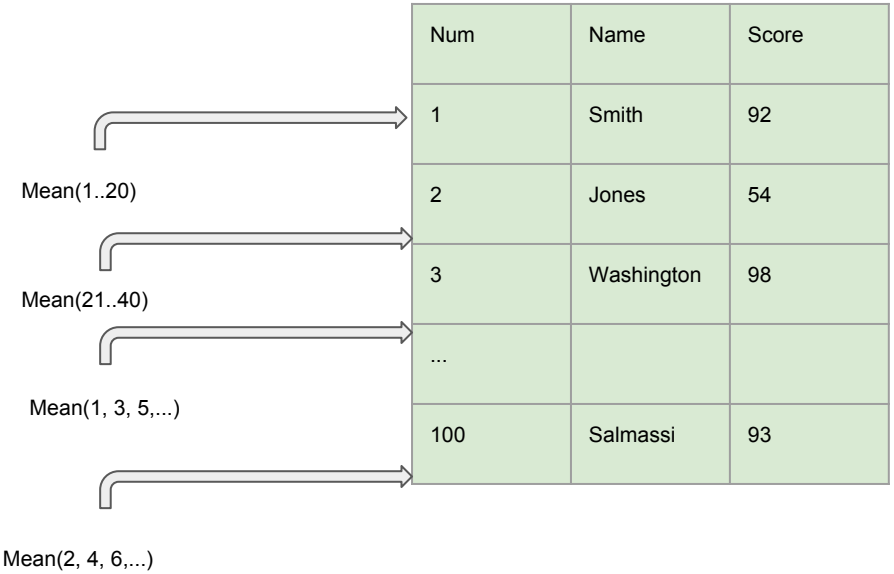
Example 2: Machine Learning Helps Identify Root Causes of Anomalous Network Traffic

- Microsoft Azure - Principal Component Analysis to classify traffic anomalies; identify security threats
  - “PCA converts a set of cases, containing possibly correlated variables, into a set of values called principal components. In the case of anomaly detection, for each new input, the anomaly detector first computes its projection on the eigenvectors, and then computes the normalized reconstruction error. This normalized error is the anomaly score. The higher the error, the more anomalous the instance is.”
    - Ref https://msdn.microsoft.com/en-us/library/azure/dn913102.aspx



Example 3: Statistical Analysis Makes Inference Attacks Easier; Harms Anonymity & Confidentiality

- Inference Attack: learning something you shouldn't know via indirect questioning
  - E.g. database of 100 people. Allowed to know aggregate properties - mean, median, etc. - of groups of least 10.
  - Inference attack: repeatedly query with different sets of subjects - mean of 1-20; mean of 21-40; mean of 2,4,6..40; etc.
  - Becomes a system of linear equations - with enough equations determining the value of any variable is solvable
- With high-powered statistical tools, implementing such attacks becomes feasible on large data sets



Example 4: Data Science has Entirely Changed Malware Detection; Helps Accountability, Integrity, Confidentiality

- Previous: signature match
- Now:
  - Classification; machine learning; predictive analytics
    - Lookout Corp.,
      - Real-time security telemetry from global population of devices
      - Sift through data to identify complex risk correlations that would otherwise evade analysis
        - https://www.lookout.com/docs/Lookout%20Security%20Platform_%20Technology%20Whitepaper.pdf
    - RiskRanker (Grace, et alia)
      - First stage analysis: signature detection/ static code analysis
      - Second stage analysis: behavioral modeling, dynamic analysis
        - Look for encrypted native code, unsafe byte code loading,...
        - http://www.facweb.iitkgp.ernet.in/~niloy/COURSE/Autumn2014/SmartPhone/Books_Paper/RiskRanker.pdf

Example 5: Data Analytics Facilitate Aggregation Attacks, Harm Anonymity & Confidentiality

- In many situations, release of a single data item is harmless; release of a large dataset is unacceptable
  - E.g., release of a single employee's contact info is okay; release of the company employee database is not
- "Aggregation attacks" involve repeated access/querying of a dataset to eventually get (a large portion of) the full dataset
- Used to involve repeated querying of a database, which could be detected/tracked/blocked
- With open data and analytics tools, it may be possible to aggregate information by getting it from multiple sources and combining them
  - E.g.: lists of employees participating in standards bodies
  - + lists of employees working on specific contracts
  - + LinkedIn profiles mentioning the company
  - + Facebook profiles mentioning the company
  - + required disclosures (e.g., SEC filings)
  - + …
  - = a respectable part of the employee database

What Should An Organization Do?

What Should an Organization Do?

- Prior to releasing data:
  - Understand what data they have, and from who
    - Subjects, researchers/contributors
  - Understand what security and privacy requirements are today
  - Understand what data science capabilities exist today
    - Analysis capabilities never shrink; they will only get stronger
  - Understand what other datasets exist today and are planned to exist
    - De-anonymization/inference/aggregation attacks often make use of multiple datasets
  - Analyze the situation themselves to identify the risks
- Federal Government:
  - Follow NIST Guidance on security
  - Implement privacy controls where feasible
- Continuous monitoring of system/data/technology/attacks

The Data Scientist's Algorithm

- First, be an advocate for sharing of data where it doesn't harm security and/or privacy
- Take responsibility for understanding what data you have and the impact of releasing it
  - Make sure Chief Privacy Officer/Senior Agency Official for Privacy understands
- Understand the state of the art in security, privacy and data science
  - Work with subject-matter experts if necessary
- Monitor technical and regulatory developments
  - New attacks
  - New defenses
  - New laws/regulations
- Be willing to adapt and react if necessary

Questions?

Different Definitions of Privacy

We used the definition from Westin (1967)

> “the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others.”

What if we used some other definition?

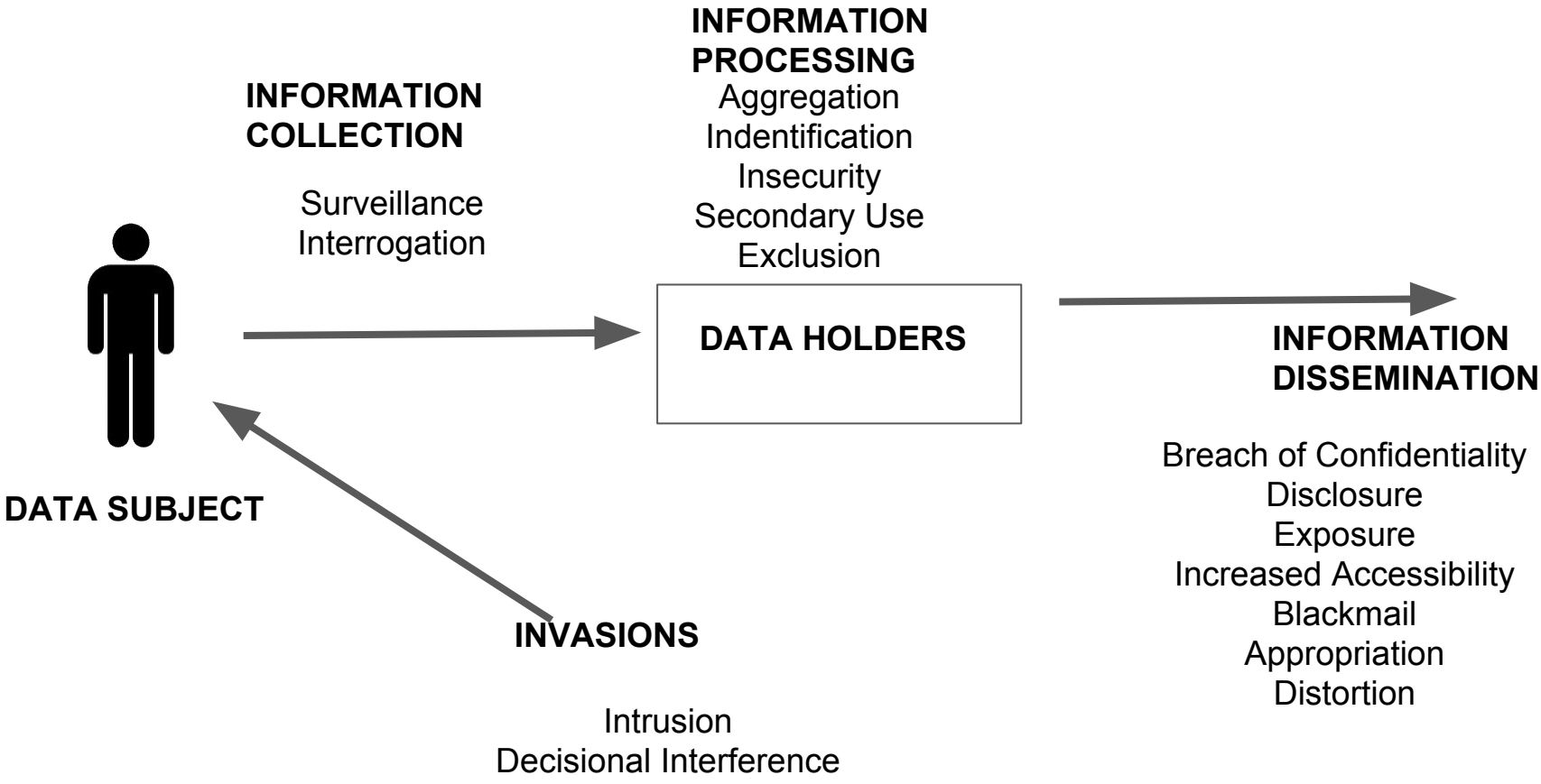
Nissenbaum - developing a single definition of ***privacy*** is a waste of time. What’s important is contextual integrity

- This is consistent with our approach. The privacy facets are relevant in the Government context

Solove - no singular essence or “core” characteristic of ***privacy***. Taxonomy that focuses more specifically on activities that invade privacy.

- We took a subset of his taxonomy. Some was already covered in Security; some was out of scope

Solove's Privacy Taxonomy



FYI: NIST SP 800-53 Privacy Controls

AP-1 AUTHORITY TO COLLECT	DM-2 DATA RETENTION AND DISPOSAL
AP-2 PURPOSE SPECIFICATION	DM-3 MINIMIZATION OF PII USED IN TESTING, TRAINING, AND RESEARCH
AR-1 GOVERNANCE AND PRIVACY PROGRAM	IP-1 CONSENT
AR-2 PRIVACY IMPACT AND RISK ASSESSMENT	IP-2 INDIVIDUAL ACCESS
AR-3 PRIVACY REQUIREMENTS FOR CONTRACTORS AND SERVICE PROVIDERS	IP-3 REDRESS
AR-4 PRIVACY MONITORING AND AUDITING	IP-4 COMPLAINT MANAGEMENT
AR-5 PRIVACY AWARENESS AND TRAINING	SE-1 INVENTORY OF PERSONALLY IDENTIFIABLE INFORMATION
AR-6 PRIVACY REPORTING	SE-2 PRIVACY INCIDENT RESPONSE
AR-7 PRIVACY-ENHANCED SYSTEM DESIGN AND DEVELOPMENT	TR-1 PRIVACY NOTICE
AR-8 ACCOUNTING OF DISCLOSURES	TR-2 SYSTEM OF RECORDS NOTICES AND PRIVACY ACT STATEMENTS
DI-1 DATA QUALITY	TR-3 DISSEMINATION OF PRIVACY PROGRAM INFORMATION
DI-2 DATA INTEGRITY AND DATA INTEGRITY BOARD	UL-1 INTERNAL USE (Limitation)
DM-1 MINIMIZATION OF PERSONALLY IDENTIFIABLE INFORMATION	UL-2 INFORMATION SHARING WITH THIRD PARTIES