

CSCSE 689: Special Topics in Modern Algorithms for Data Science

Lecture 32

Samson Zhou

Presentation Schedule

- **November 27:** Chunkai, Jung, Galaxy AI
- **November 29:** STMI, Anmol, Jason
- **December 1:** Bokun, Ayesha, Dawei, Lipai

census.gov:



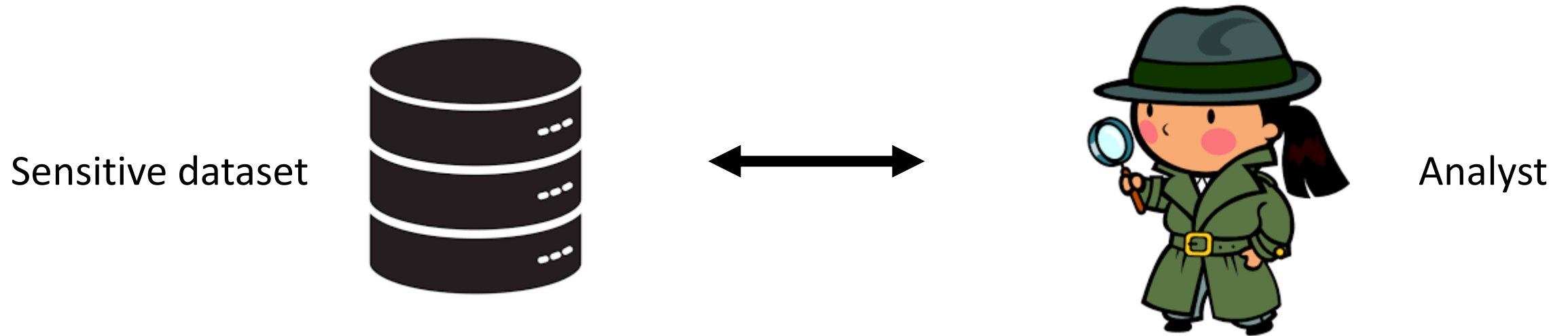
Privacy & Confidentiality

Federal Law Protects Your Information. The U.S. Census Bureau is bound by [Title 13](#) of the United States Code. This law not only provides authority for the work we do, but also provides strong protection for the information we collect from individuals and businesses. As a result, the Census Bureau has one of the strongest confidentiality guarantees in the federal government.

It is against the law for any Census Bureau employee to disclose or publish any census or survey information that identifies an individual or business. This is true even for inter-agency communication: the FBI and other government entities do not have the legal right to access this information. In fact, when these protections have been challenged, Title 13's confidentiality guarantee has been upheld.

For more information about how the Census Bureau safeguards the data it collects, visit the agency's [Data Protection](#) and [Disclosure Avoidance Working Papers](#) Web sites.

Private Data Analysis



- Analysis of medical datasets to predict possible issues
- Pattern detection for social networks or epidemic spread
- US Census information for apportionment

Anonymization

Sensitive dataset



Anonymized dataset



Analyst

Anonymizing Data

Age	Zip Code	Employer	Has Pet
56	77005	Apple	Yes
32	77005	Microsoft	No
71	77005	Amazon	Yes
44	77005	Petsmart	Yes
25	77005	Netflix	No
61	77005	Google	No

Anonymizing Data

Age	Zip Code	Employer	Has Pet
56	77005	Apple	Yes
32	77005	Microsoft	No
71	77005	Amazon	Yes
44	77005	Petsmart	Yes
25	77005	Netflix	No
61	77005	Google	No

Name	Age	Gender	Employer
Alice	56	Female	Apple
Bob	32	Male	Microsoft
Carol	71	Female	Amazon
Dale	44	Male	Petsmart
Erin	25	Female	Netflix
Fred	61	Male	Google


Reconstruction Attack

Name	Age	Zip Code	Gender	Employer	Has Pet
Alice	56	77005	Female	Apple	Yes
Bob	32	77005	Male	Microsoft	No
Carol	71	77005	Female	Amazon	Yes
Dale	44	77005	Male	Petsmart	Yes
Erin	25	77005	Female	Netflix	No
Fred	61	77005	Male	Google	No

Anonymizing Data

The New York Times

Netflix Cancels Contest After Concerns Are Raised About Privacy

 Share full article



By **Steve Lohr**

March 12, 2010

👍		👎	👍		
	👍				
👍		👎		👍	👍
👍			👎		
	👍		👎	👎	
		👎	👍		

+

👍			👍		
	👍				
👍					👍
👍			👎		
				👎	
		👎			

- Alice
- Bob
- Charlie
- Danielle
- Erica
- Frank

Anonymized
NetFlix data

Public, incomplete
IMDB data

=

👍		👎	👍		
	👍				
👍		👎		👍	👍
👍			👎		
	👍		👎	👎	
		👎	👍		

- Alice
- Bob
- Charlie
- Danielle
- Erica
- Frank

Identified NetFlix Data

Differencing Attacks

- How many people in this classroom went to Kyle Field last weekend?
- How many people in this classroom besides the instructor went to Kyle Field last weekend?

US Census Bureau

Work Area Profile Analysis
enter your own subtitle

▼ Display Settings
 Characteristic Filter Total
 Year 2017

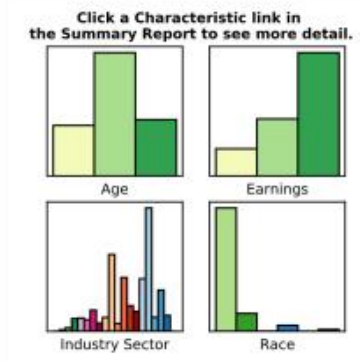
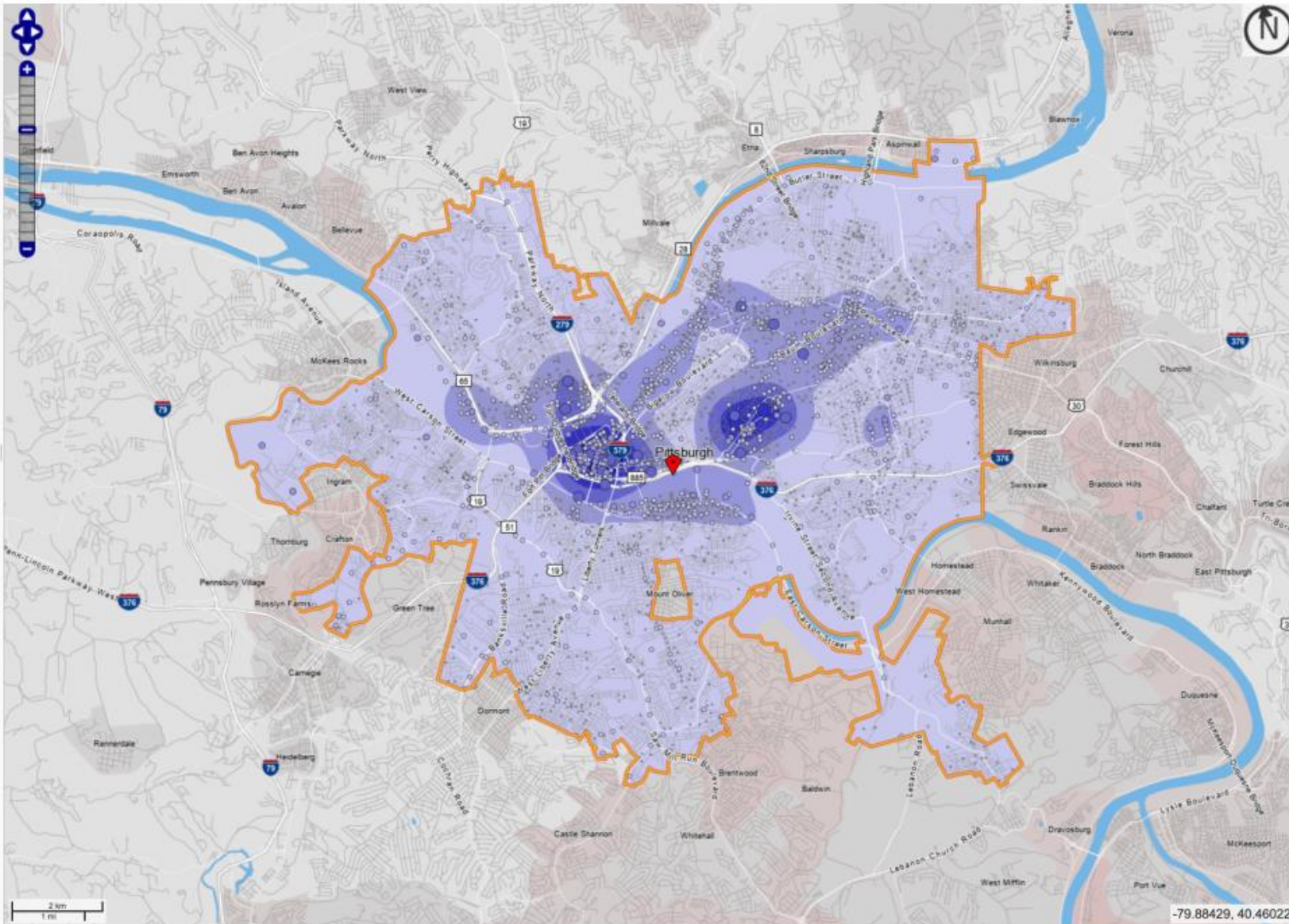
▼ Map Controls
 Color Key
 Thermal Overlay
 Point Overlay
 Selection Outline
 Identify Zoom to Selection
 Clear Overlays Animate Overlays

▼ Report/Map Outputs
 Detailed Report
 Export Geography
 Print Chart/Map

▼ Legends
 5 - 4,815 Jobs/Sq.Mile
 4,816 - 19,245 Jobs/Sq.Mile
 19,246 - 43,295 Jobs/Sq.Mile
 43,296 - 76,965 Jobs/Sq.Mile
 76,966 - 120,256 Jobs/Sq.Mile
 - 1 - 26 Jobs
 - 27 - 416 Jobs
 - 417 - 2,103 Jobs
 - 2,104 - 6,645 Jobs
 - 6,646 - 16,223 Jobs
 Analysis Selection

▼ Analysis Settings

Change Settings



View as Bar Chart

Total Private Primary Jobs

	2017	Count	Share
Total Private Primary Jobs		246,264	100.0%

Worker Age

	2017	Count	Share
Age 29 or younger		54,080	22.0%
Age 30 to 54		131,951	53.6%
Age 55 or older		60,233	24.5%

Earnings

	2017	Count	Share
\$1,250 per month or less		32,261	13.1%
\$1,251 to \$3,333 per month		67,721	27.5%
More than \$3,333 per month		146,282	59.4%

NAICS Industry Sector

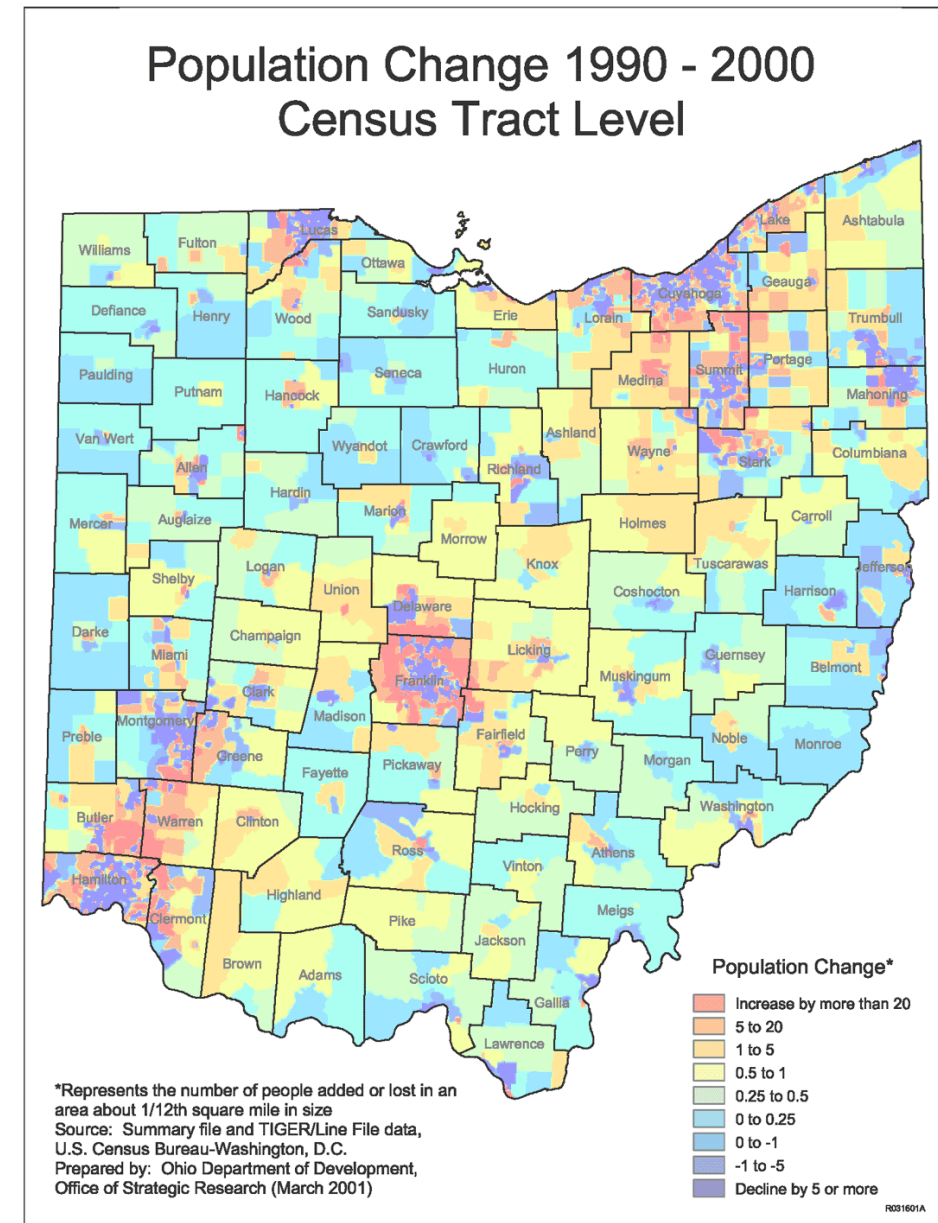
	2017	Count	Share
Agriculture, Forestry, Fishing and Hunting		2	0.0%
Mining, Quarrying, and Oil and Gas Extraction		709	0.3%
Utilities		1,742	0.7%
Construction		6,057	2.5%
Manufacturing		6,311	2.6%
Wholesale Trade		5,550	2.3%

2010 US Census

- **308,745,738** people \times **6** variables = **1,852,473,228** measurements collected
- Total statistics: **5,578,897,932**
- Create a system of **5.5** billion equations with **1.8** billion unknowns

2010 US Census

- Reconstruction attack on 2010 US Census by researchers recovered information for **308,745,538** people using census block and tract summary tables



Summary

- “Ad-hoc” privacy procedures like anonymization/deidentification often fails
- Publishing too many queries on a sensitive database with too much accuracy can compromise the privacy of the database
- Need a formal mathematical notion for measuring privacy

Possible Notion for Privacy #1

- “The data analyst cannot learn anything about Alice”



Sensitive dataset



Analyst

Possible Notion for Privacy #1

- “The data analyst cannot learn anything about Alice”



Alice is known to
be an Aggie



Sensitive dataset



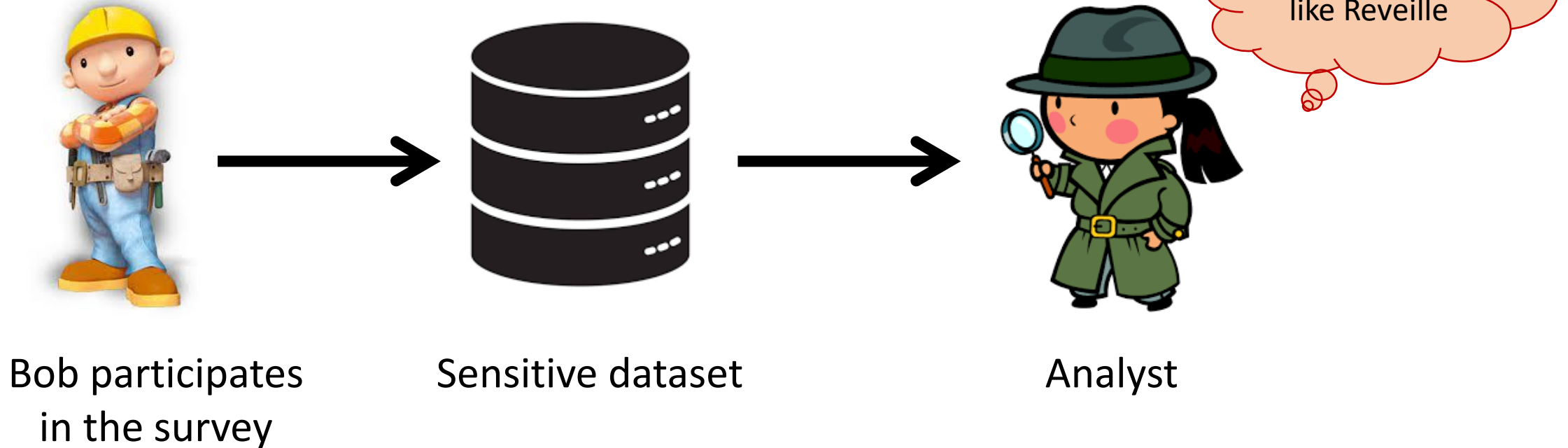
Analyst



Was Alice's privacy violated?

Possible Notion for Privacy #1

- “The data analyst cannot learn anything about Alice”



Even though Alice is not in the survey, it is still known that Alice is an Aggie

Possible Notion for Privacy #1

- Suppose a survey is conducted on a sensitive dataset and concludes that *“most Aggies like dogs, e.g., Reveille”*
- Alice is a known Aggie, and so a data analyst infers that Alice is more likely to be a dog owner and asks for higher apartment cleaning rates
- **Was Alice’s privacy violated by this study?**



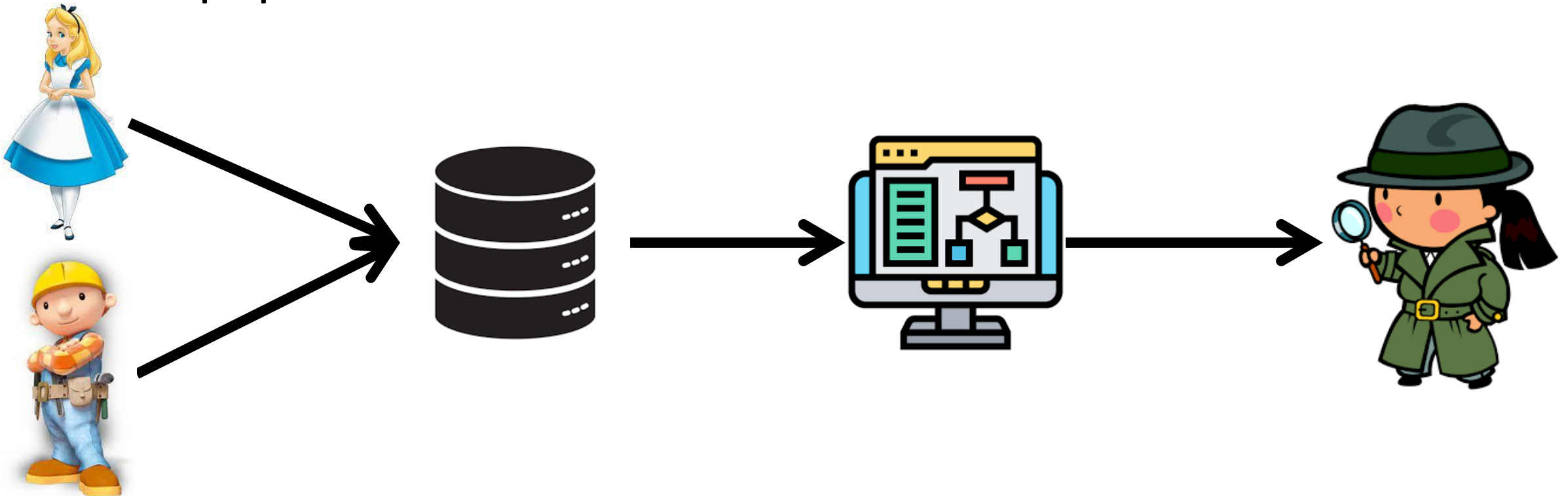
Possible Notion for Privacy #2

- “A study is private...if the data analyst gains *almost no additional information* about Alice from the study than if the same study was performed *without Alice’s data*”



Possible Notion for Privacy #2

- **Stability**: the data analyst reaches roughly similar conclusions if any individual data point is replaced by another data point of the population

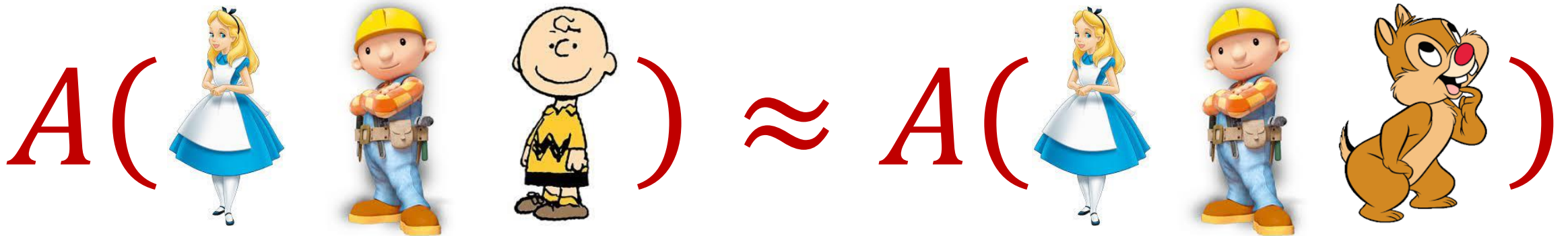


Differential Privacy

- [DMNS06] Given $\varepsilon > 0$ and $\delta \in (0,1)$, a randomized algorithm $A: U^* \rightarrow Y$ is (ε, δ) -differentially private if, for every neighboring frequency vectors f and f' and for all $E \subseteq Y$,
$$\Pr[A(f) \in E] \leq e^\varepsilon \cdot \Pr[A(f') \in E] + \delta$$

Differential Privacy

- [DMNS06] Given $\epsilon > 0$ and $\delta \in (0,1)$, a randomized algorithm $A: U^* \rightarrow Y$ is (ϵ, δ) -differentially private if, for every neighboring frequency vectors f and f' and for all $E \subseteq Y$,
$$\Pr[A(f) \in E] \leq e^\epsilon \cdot \Pr[A(f') \in E] + \delta$$



Differential Privacy

- [DMNS06] Given $\varepsilon > 0$ and $\delta \in (0,1)$, a randomized algorithm $A: U^* \rightarrow Y$ is (ε, δ) -differentially private if, for every neighboring frequency vectors f and f' and for all $E \subseteq Y$,
$$\Pr[A(f) \in E] \leq e^\varepsilon \cdot \Pr[A(f') \in E] + \delta$$

- For small ε , can think of e^ε as $1 + \varepsilon$

$$\Pr[A(f) \in E] \leq (1 + \varepsilon) \cdot \Pr[A(f') \in E] + \delta$$

Differential Privacy

- [DMNS06] Given $\epsilon > 0$ and $\delta \in (0,1)$, a randomized algorithm $A: U^* \rightarrow Y$ is (ϵ, δ) -differentially private if, for every neighboring frequency vectors f and f' and for all $E \subseteq Y$,
$$\Pr[A(f) \in E] \leq e^\epsilon \cdot \Pr[A(f') \in E] + \delta$$
- δ can be interpreted as the probability that the mechanism “fails” to be differentially private

Differential Privacy

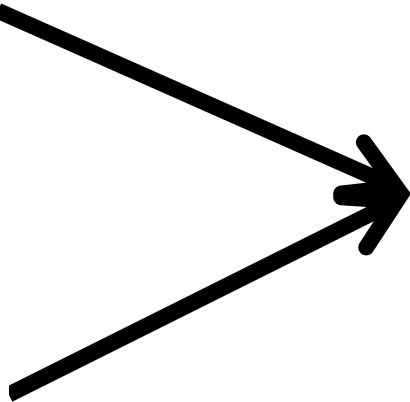
- [DMNS06] Given $\epsilon > 0$ and $\delta \in (0,1)$, a randomized algorithm $A: U^* \rightarrow Y$ is (ϵ, δ) -differentially private if, for every neighboring frequency vectors f and f' and for all $E \subseteq Y$,
$$\Pr[A(f) \in E] \leq e^\epsilon \cdot \Pr[A(f') \in E] + \delta$$

- If $\delta = 0$, a mechanism is said to satisfy *pure differential privacy*
- Otherwise if $\delta > 0$, a mechanism is said to satisfy *approximate differential privacy*

Differential Privacy

- [DMNS06] Given $\epsilon > 0$ and $\delta \in (0,1)$, a randomized algorithm $A: U^* \rightarrow Y$ is (ϵ, δ) -differentially private if, for every neighboring frequency vectors f and f' and for all $E \subseteq Y$,

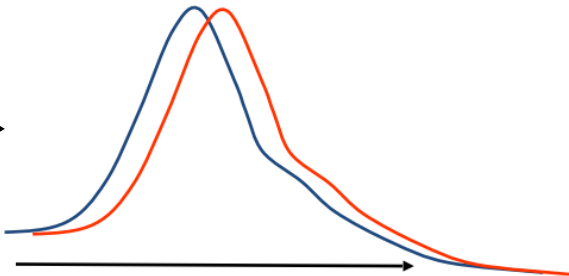
$$\Pr[A(f) \in E] \leq e^\epsilon \cdot \Pr[A(f') \in E] + \delta$$



Sensitive dataset



Algorithm



Output distribution

Differential Privacy

- [DMNS06] Given $\varepsilon > 0$ and $\delta \in (0,1)$, a randomized algorithm $A: U^* \rightarrow Y$ is (ε, δ) -differentially private if, for every neighboring frequency vectors f and f' and for all $E \subseteq Y$,
$$\Pr[A(f) \in E] \leq e^\varepsilon \cdot \Pr[A(f') \in E] + \delta$$
- **Implication:** Deterministic algorithms cannot be differentially private unless they are a constant function