Pan-Private Uniformity Testing

Matthew Joseph

Kareem Amin

Jieming Mao

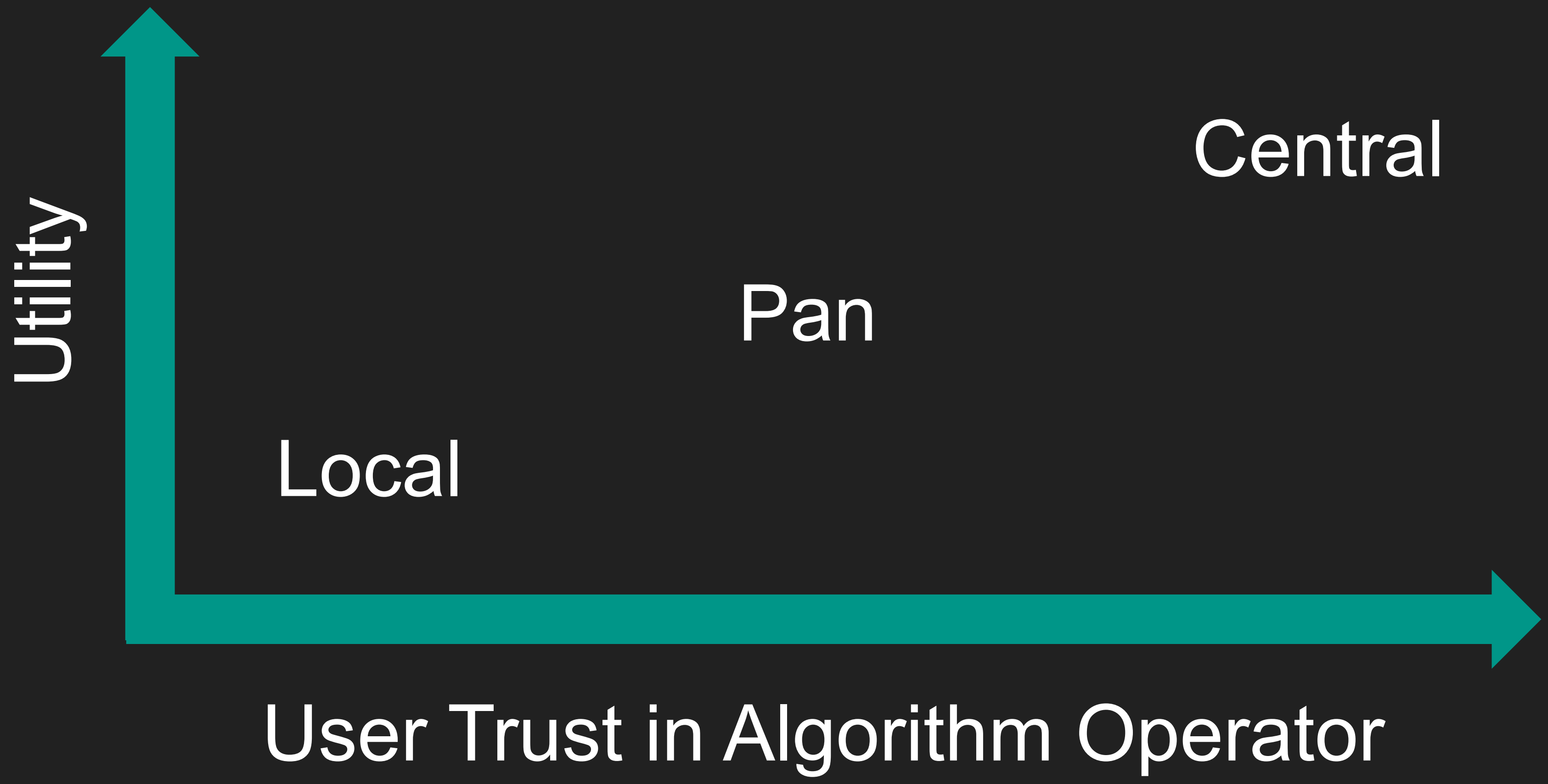
Models of Differential Privacy

Central

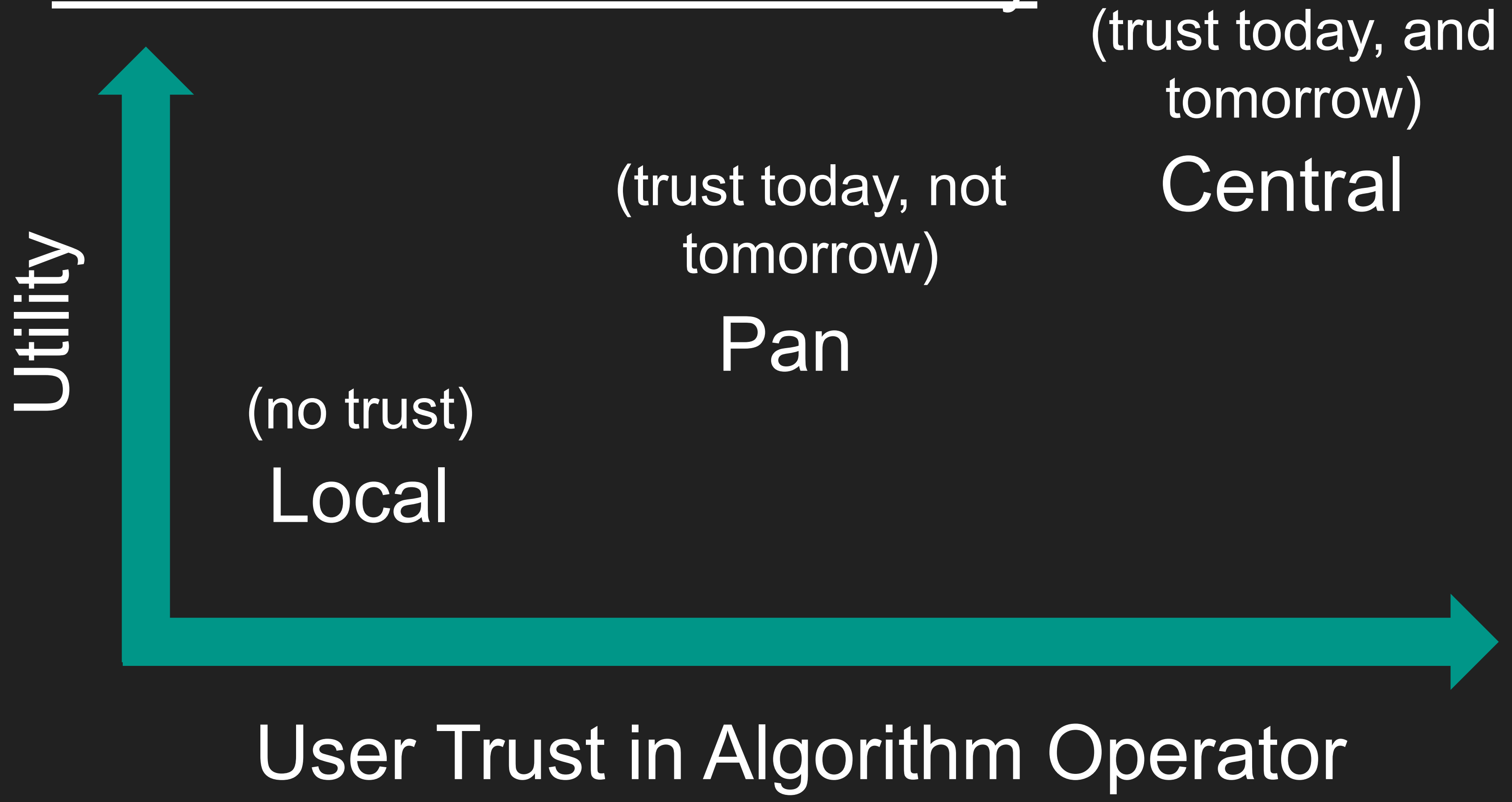
Pan

Local

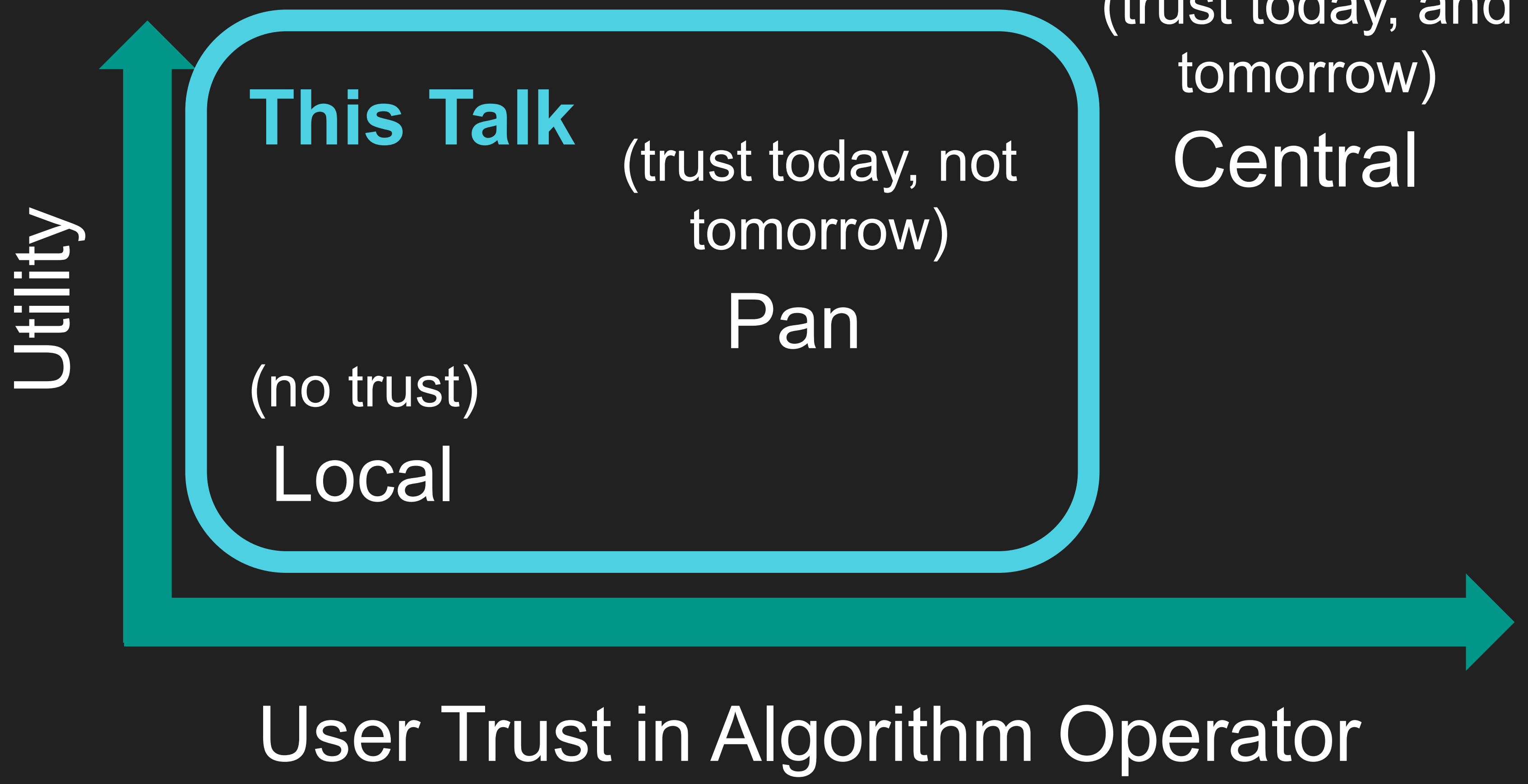
Models of Differential Privacy



Models of Differential Privacy



Models of Differential Privacy



Outline

1. Local Privacy Basics
2. Pan-Privacy Basics
3. Result 1: Connecting Local and Pan-Privacy
4. Result 2: Pan-Private Uniformity Testing

Outline

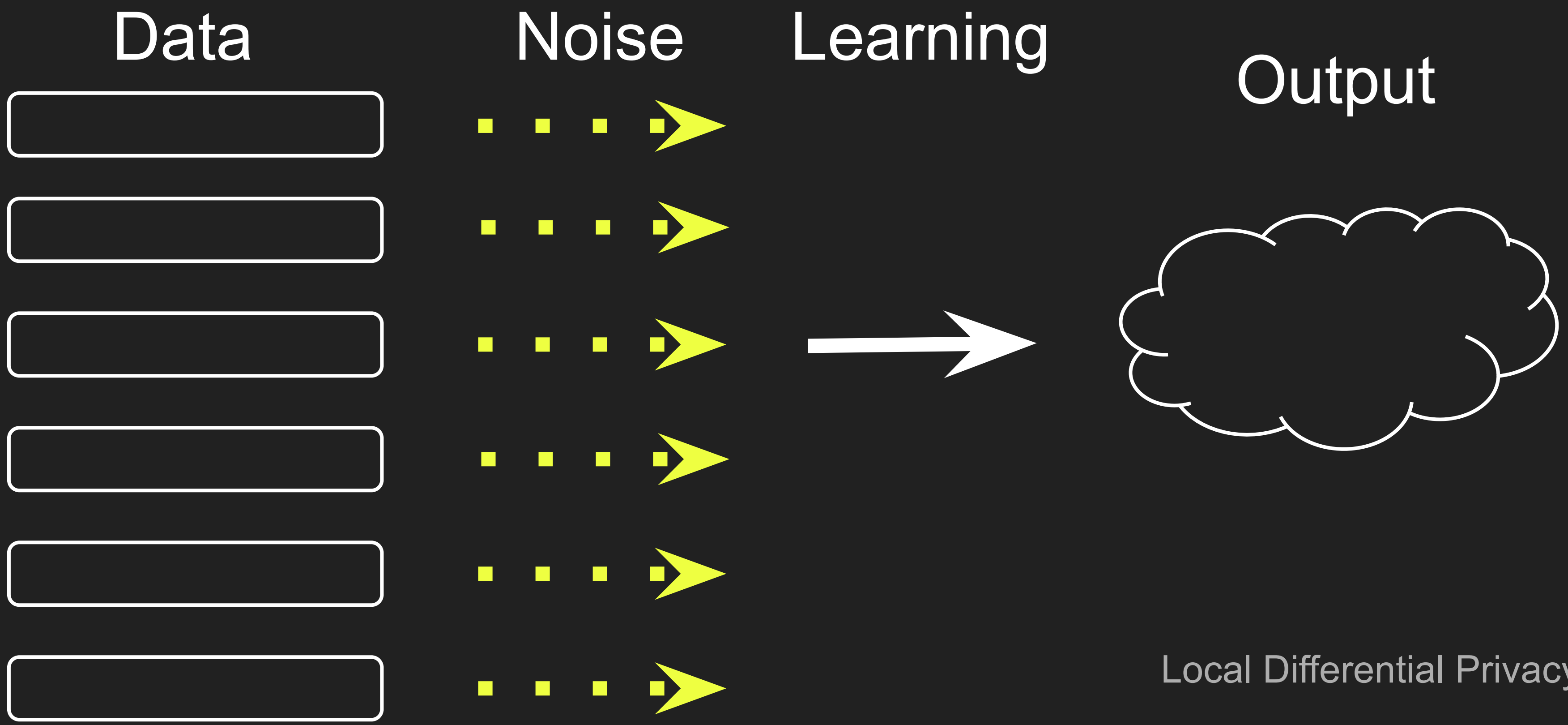
1. **Local Privacy Basics**

2. Pan-Privacy Basics

3. Result 1: Connecting Local and Pan-Privacy

4. Result 2: Pan-Private Uniformity Testing

Local DP Learning From Data



Local Differential Privacy

Local DP in Words

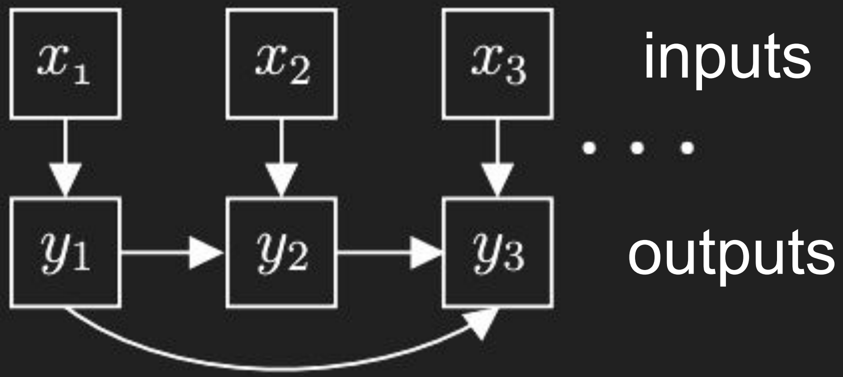
Distributed database, users keep their data

Protocol $\boldsymbol{A}$ learns about the data through public communication with users

Users send responses through *randomizers* $\boldsymbol{R}$, differentially private functions of one datum

Types of LDP Interactivity

Definition: Protocol A is *sequentially interactive* [DJW13] if all users speak once (possibly in multiple rounds).



Local DP in Math

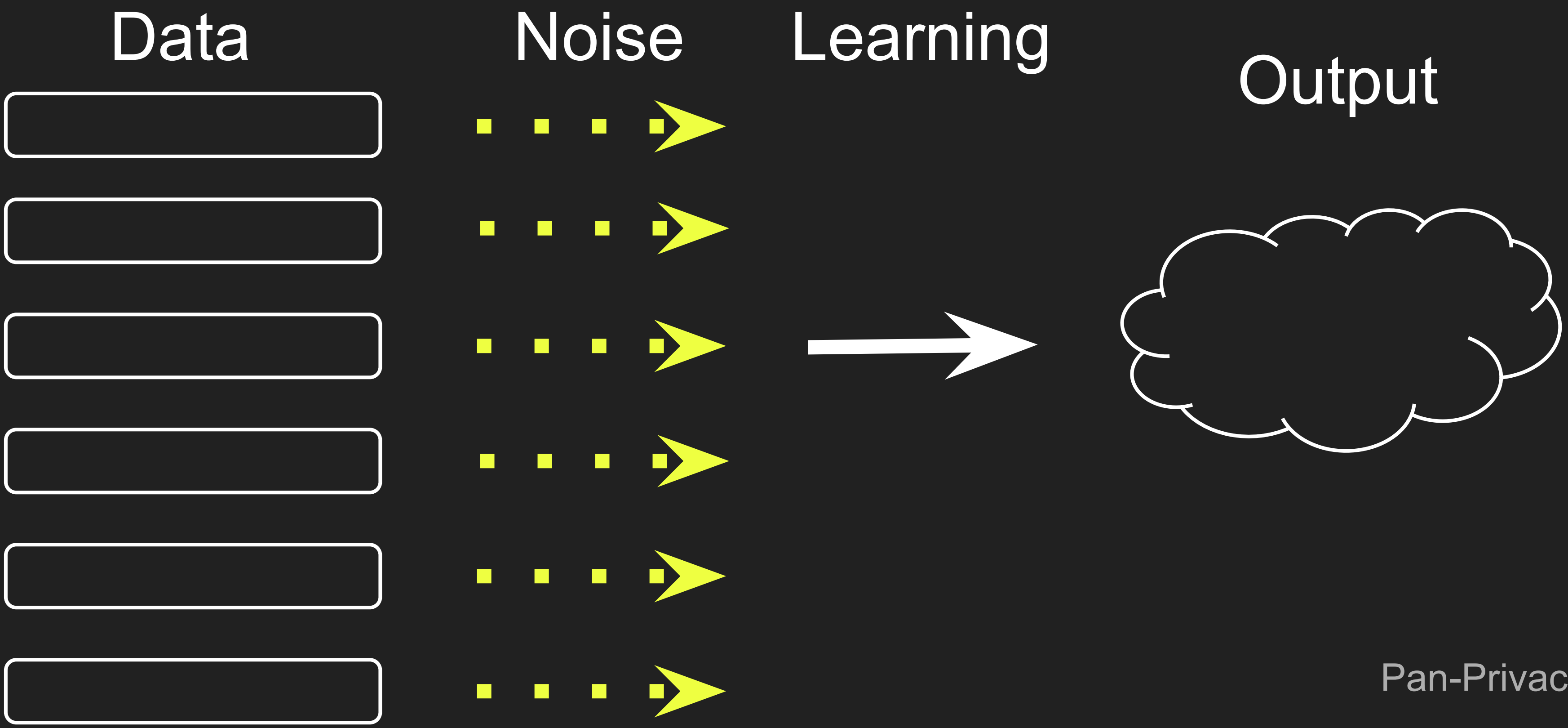
Definition: Sequentially interactive protocol $\mathcal{A}$ is (ε,δ)-locally differentially private (LDP) if all randomizers are (ε,δ)-randomizers.

$$(P[R(x) \text{ in } Y] \leq e^{\varepsilon} P[R(x') \text{ in } Y] + \delta)$$

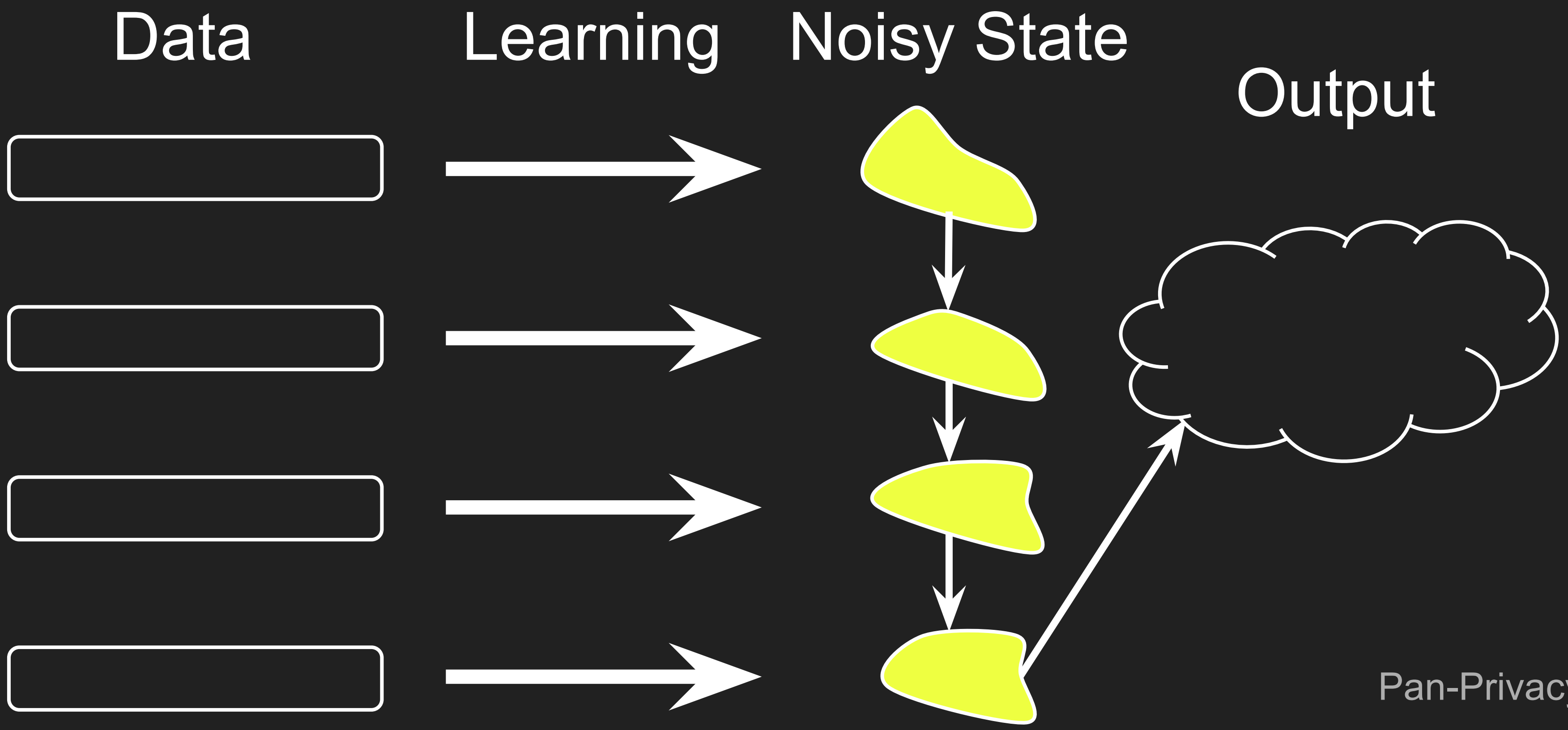
Outline

1. Local Privacy Basics
2. **Pan-Privacy Basics**
3. Result 1: Connecting Local and Pan-Privacy
4. Result 2: Pan-Private Uniformity Testing

Local DP Learning From Data



Pan-Private [DNPRY10] Learning From Data



Pan-Privacy in Words

Data arrives in a stream, one element at a time

Algorithm $\mathcal{A}$ sees element, updates internal state, continues

Adversary sees (any) one internal state and final output, and this view must be a differentially private function of the stream

See data (easier than local), private intermediary state (harder than central)

Pan-Privacy in Math

Definition: Streams S and S' are neighbors if they differ in at most one stream element. Protocol A is (ε,δ)-pan private against one intrusion if, for all neighboring S and S', times t, internal state subsets $\mathcal{I}$, and output subsets $\mathcal{O}$,

$$P[I(S_{\leq t}) \text{ in } \mathcal{I}, O(S_{\leq t} \circ S_{>t}) \text{ in } \mathcal{O}] \leq$$

$$e^{\varepsilon} P[I(S_{\leq t}') \text{ in } \mathcal{I}, O(S_{\leq t}' \circ S_{>t}') \text{ in } \mathcal{O}] + \delta.$$

Why Pan-Privacy?

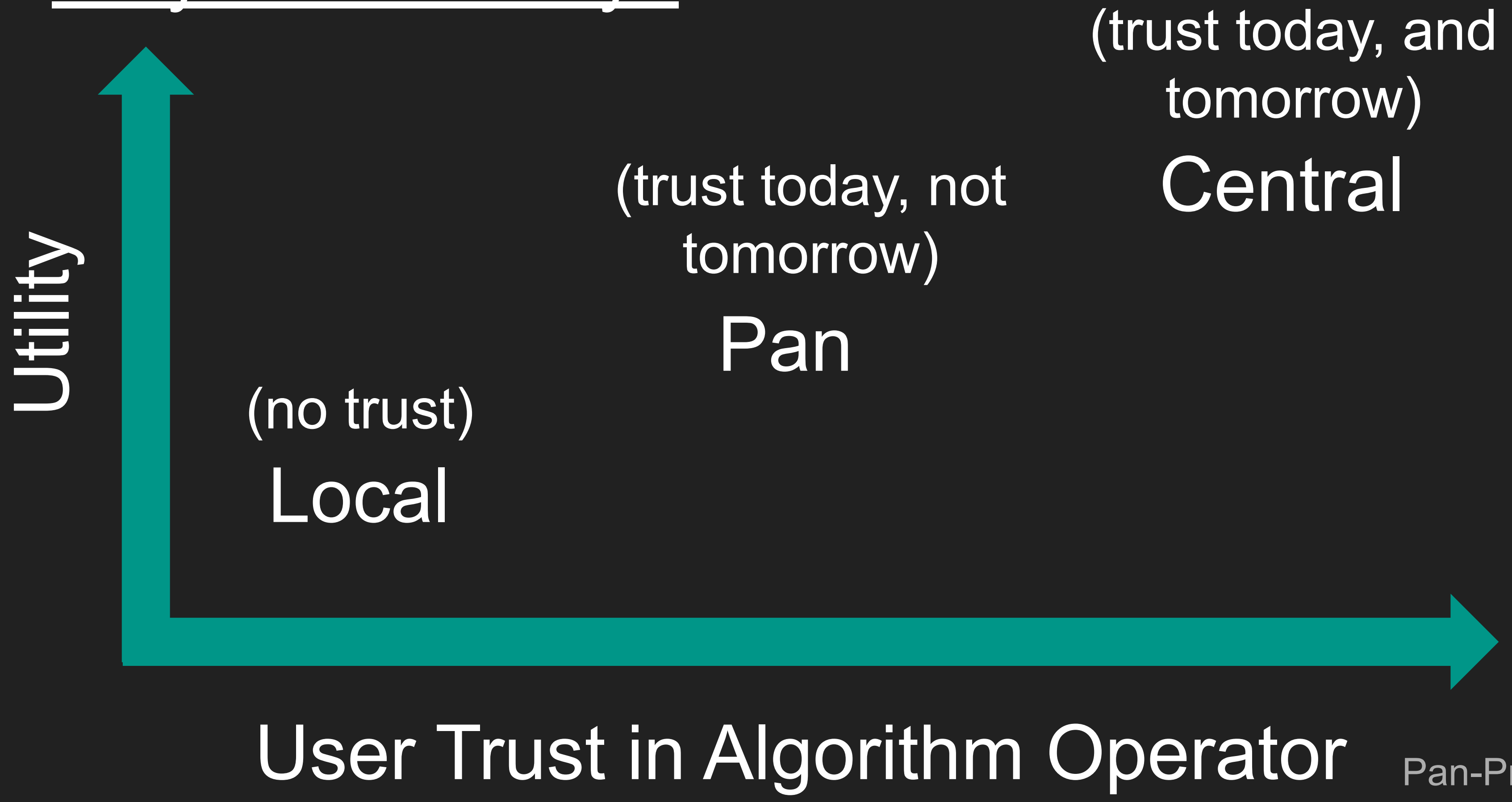
Why Pan-Privacy?

Most useful when user trusts operator today, but wants to “future-proof” their data

Examples: worried about government subpoena or operator ownership changes

If user trusts the operator today, privacy of intermediate state protects against future intrusions

Why Pan-Privacy?



Q: Does the one-intrusion assumption matter?

Q: Does the one-intrusion assumption matter?

A: Yes

Outline

1. Local Privacy Basics
2. Pan-Privacy Basics
3. **Result 1: Connecting Local and Pan-Privacy**
4. Result 2: Pan-Private Uniformity Testing

Result 1: Pan- vs. Local

Theorem: Any algorithm A_P that is ε-pan-private against two intrusions can be converted into an identical sequentially interactive ε-LDP protocol A_S, and vice-versa.

Result 1: Pan- vs. Local

Theorem: Any algorithm A_P that is ε-pan-private against two intrusions can be converted into an identical sequentially interactive ε-LDP protocol A_S, and vice-versa.

So if you need privacy against multiple intrusions, may as well use local privacy.

Result 1: Pan- vs. Local

Proof Sketch

Local to pan: run a local protocol and maintain transcript as internal state.

Result 1: Pan- vs. Local

Proof Sketch

Local to pan: run a local protocol and maintain transcript as internal state.

Pan to local: adversary sees two internal states, can “diff” them. So must randomize whenever update internal state. Randomize every state ≈ sequential interactivity.

Q: Is single-intrusion pan-privacy meaningful?

Q: Is single-intrusion pan-privacy meaningful?

A: We suggest yes

Why Single-Intrusion Pan-Privacy?

Single-intrusion pan-privacy suffers when a user contributes data between intrusions (“diff” attack)

Users most worried about giving data to an operator that’s already compromised

For users who trust operator today, single-intrusion pan-privacy is useful (and more private than central)

Outline

1. Local Privacy Basics
2. Pan-Privacy Basics
3. Result 1: Connecting Local and Pan-Privacy
4. **Result 2: Pan-Private Uniformity Testing**

Result 2: Pan-Private Uniformity Testing

Uniformity testing: algorithm receives samples from unknown distribution p over $[k]$ and must distinguish $p = U_k$ from

$||p - U_k||_{TV} \geq \alpha$ w.p. $\geq 2/3$

Result 2: Pan- Uniformity Testing

	Previous Work	This Work
Without Privacy	$\Theta(k^{1/2})$ [CDVV14]	
ε-DP	$\Theta(k^{1/2})$ [ASZ18]	
ε-Pan Privacy		
SI ε-LDP		
NI ε-LDP	$\Theta(k)$ [ACFT19]	

Result 2: Pan- Uniformity Testing

	Previous Work	This Work
Without Privacy	$\Theta(k^{1/2})$ [CDVV14]	
ε-DP	$\Theta(k^{1/2})$ [ASZ18]	
ε-Pan Privacy		$\Theta(k^{2/3})$
SI ε-LDP		
NI ε-LDP	$\Theta(k)$ [ACFT19]	

Result 2: Pan- Uniformity Testing

	Previous Work	This Work
Without Privacy	$\Theta(k^{1/2})$ [CDVV14]	
ε-DP	$\Theta(k^{1/2})$ [ASZ18]	
ε-Pan Privacy		$\Theta(k^{2/3})$
SI ε-LDP		$\Theta(k)$
NI ε-LDP	$\Theta(k)$ [ACFT19]	

Result 2: Pan- Uniformity Testing

Theorem: ε-pan-private uniformity testing has sample complexity

$$O\left(\frac{k^{2/3}}{\alpha^{4/3}\varepsilon^{2/3}} + \frac{\sqrt{k}}{\alpha^2} + \frac{\sqrt{k}}{\alpha\varepsilon}\right)$$

$$\Omega\left(\frac{k^{2/3}}{\alpha^{4/3}\varepsilon^{2/3}} + \frac{\sqrt{k}}{\alpha^2} + \frac{\sqrt{k}}{\alpha\sqrt{\varepsilon}} + \frac{1}{\alpha\varepsilon}\right)$$

Result 2: Pan- Uniformity Testing

Upper Bound Sketch

Result 2: Pan- Uniformity Testing

Upper Bound Sketch

Key idea: split difference between central and local approaches

Result 2: Pan- Uniformity Testing

Upper Bound Sketch

Central [CDK17, ADR18, ASZ18]: uses “fine” statistic

Looks at sample counts for all k elements and measures departure from expected count under uniform distribution

Need to add noise to each count to be pan-private

Result 2: Pan- Uniformity Testing

Upper Bound Sketch

Central [CDK17, ADR18, ASZ18]: uses “fine” statistic

Looks at sample counts for all k elements and measures departure from expected count under uniform distribution

Need to add noise to each count to be pan-private. Can get pan- $\mathbf{O}(k^{3/4})$ like this...but can we do better?

Result 2: Pan- Uniformity Testing

Upper Bound Sketch

Central [CDK17, ADR18, ASZ18]: uses “fine” statistic

Maybe pan- should use a coarser statistic?

Result 2: Pan- Uniformity Testing

Upper Bound Sketch

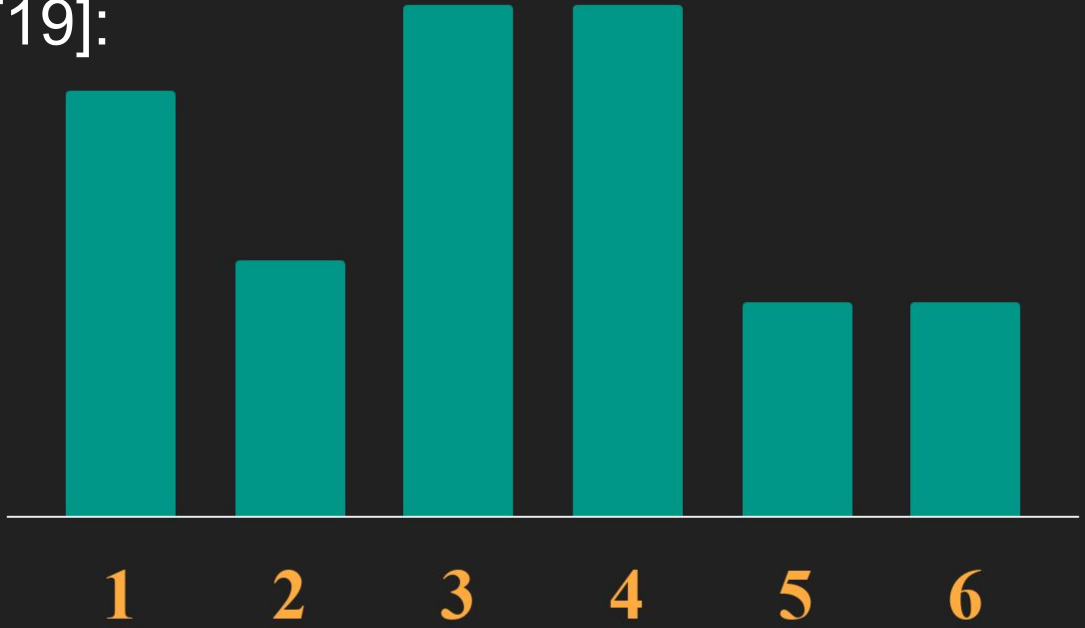
Local [ACFT19]: uses coarse statistic

Randomly halves domain, now uniformity testing over **[2]**

Result 2: Pan- Uniformity Testing

Upper Bound Sketch

Local [ACFT19]:



Result 2: Pan- Uniformity Testing

Upper Bound Sketch

Local [ACFT19]:

Result 2: Pan- Uniformity Testing

Upper Bound Sketch

Local [ACFT19]: uses coarse statistic

Small response domain: good for local!

But sacrifices a lot of testing distance: α to $\alpha/k^{1/2}$… so end up using $\mathbf{O}(k)$ samples

Result 2: Pan- Uniformity Testing

Upper Bound Sketch

Local [ACFT19]: uses coarse statistic

Maybe pan- should maintain a finer statistic?

Result 2: Pan- Uniformity Testing

Upper Bound Sketch

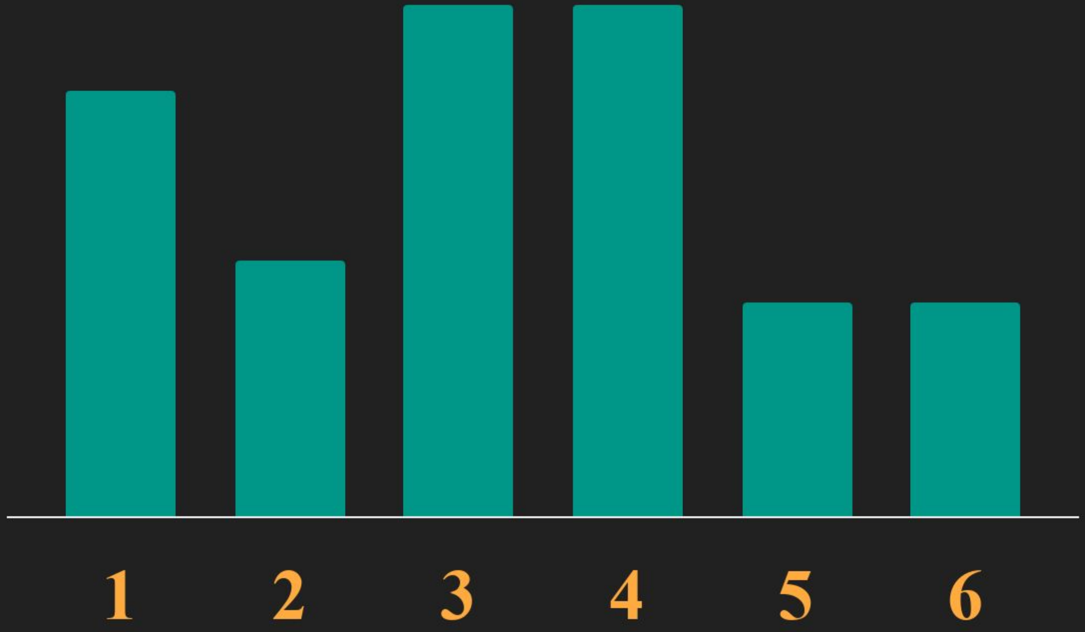
Pan: coarser than central, finer than local

Randomly partition domain into n equal-size groups, now uniformity testing over $[n]$

Result 2: Pan- Uniformity Testing

Upper Bound Sketch

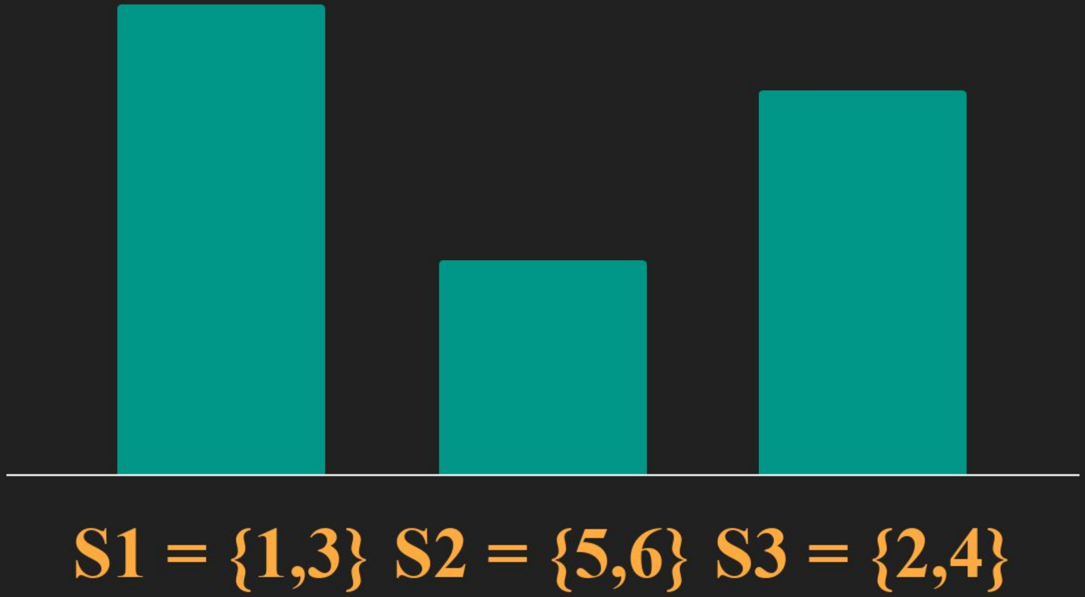
Pan:



Result 2: Pan- Uniformity Testing

Upper Bound Sketch

Pan:



Result 2: Pan- Uniformity Testing

Upper Bound Sketch

Pan: coarser than central, finer than local

Testing distance change is α to $\alpha(n/k)^{1/2}$

Pick $n = \Theta(k^{2/3}\varepsilon^{4/3}/\alpha^{4/3})$ to trade off coarse (not too much noise per bin) and fine (preserve testing distance)

Result 2: Pan- Uniformity Testing

Theorem: **ε**-pan-private uniformity testing has sample complexity

$$O\left(\frac{k^{2/3}}{\alpha^{4/3}\varepsilon^{2/3}} + \frac{\sqrt{k}}{\alpha^2} + \frac{\sqrt{k}}{\alpha\varepsilon}\right)$$

$$\Omega\left(\frac{k^{2/3}}{\alpha^{4/3}\varepsilon^{2/3}} + \frac{\sqrt{k}}{\alpha^2} + \frac{\sqrt{k}}{\alpha\sqrt{\varepsilon}} + \frac{1}{\alpha\varepsilon}\right)$$

Result 2: Pan- Uniformity Testing

Lower Bound Sketch

Adapts information theory lower bound from [DGKR19] for uniformity testing under memory restrictions

Result 2: Pan- Uniformity Testing

Lower Bound Sketch

Adapts information theory lower bound from [DGKR19] for uniformity testing under memory restrictions

Main contribution: replacing memory restriction with privacy restriction

Result 2: Pan- Uniformity Testing

Theorem: **ε**-pan-private uniformity testing has sample complexity

$$O\left(\frac{k^{2/3}}{\alpha^{4/3}\varepsilon^{2/3}} + \frac{\sqrt{k}}{\alpha^2} + \frac{\sqrt{k}}{\alpha\varepsilon}\right)$$

$$\Omega\left(\frac{k^{2/3}}{\alpha^{4/3}\varepsilon^{2/3}} + \frac{\sqrt{k}}{\alpha^2} + \frac{\sqrt{k}}{\alpha\sqrt{\varepsilon}} + \frac{1}{\alpha\varepsilon}\right)$$

Takeaways

- Pan-privacy is appropriate when user trusts algorithm operator today but maybe not tomorrow

Takeaways

- Pan-privacy is appropriate when user trusts algorithm operator today but maybe not tomorrow

- Pan-privacy against more than one intrusion is equivalent to sequentially interactive local privacy

Takeaways

- Pan-privacy is appropriate when user trusts algorithm operator today but maybe not tomorrow
- Pan-privacy against more than one intrusion is equivalent to sequentially interactive local privacy
- Pan-privacy against a single intrusion trades off both utility and privacy between central and local models
    - $\Theta(k^{1/2})$, $\Theta(k^{2/3})$, and $\Theta(k)$ uniformity testing bounds

Open Questions

- Uniformity testing:
    - close gap between pan upper and lower bounds
    - fully interactive locally private lower bound?

Open Questions

- Uniformity testing:
  - close gap between pan upper and lower bounds
  - fully interactive locally private lower bound?

- What about **(ε,δ)**-pan-privacy?

Open Questions

- Uniformity testing:
  - close gap between pan upper and lower bounds
  - fully interactive locally private lower bound?

- What about (ε,δ)-pan-privacy?

- How powerful is pan-privacy in general?

References

1. [ACFT19] “Test Without Trust: Optimal Locally Private Distribution Testing”. Acharya, Canonne, Freitag, Tyagi. AISTATS.
2. [ADR18] “Differentially Private Identity and Equivalence Testing of Discrete Distributions”. Aliakbarpour, Diakonikolas, Rubinfeld. ICML.
3. [ASZ18] “Differentially Private Identity Testing and Closeness of Discrete Distributions”. Acharya, Sun, Zhang. ICML.
4. [CDK17] “Priv’IT: Private and Sample Efficient Identity Testing”. Cai, Daskalakis, Kamath. ICML.
5. [CDVV14] “Optimal Algorithms for Testing Closeness of Discrete Distributions”. Chan, Diakonikolas, Valiant, Valiant. SODA.
6. [DGKR19] “Communication and Memory Efficient Testing of Discrete Distributions”. Diakonikolas, Gouleakis, Kane, Rao. COLT.
7. [DJW13] “Local Privacy, Data Processing Inequalities, and Statistical Minimax Rates”. Duchi, Jordan, Wainwright. FOCS.
8. [DMNS06] “Calibrating Noise to Sensitivity in Private Data Analysis”. Dwork, Mcsherry, Nissim, Smith. TCC.
9. [W65] "Randomised Response: A Survey Technique for Eliminating Evasive Answer Bias". Warner. JASA.