Localish
Distributed Models for Statistical Data Privacy

Adam Smith

BU Computer Science

PPML 2018 Workshop
December 8, 2018

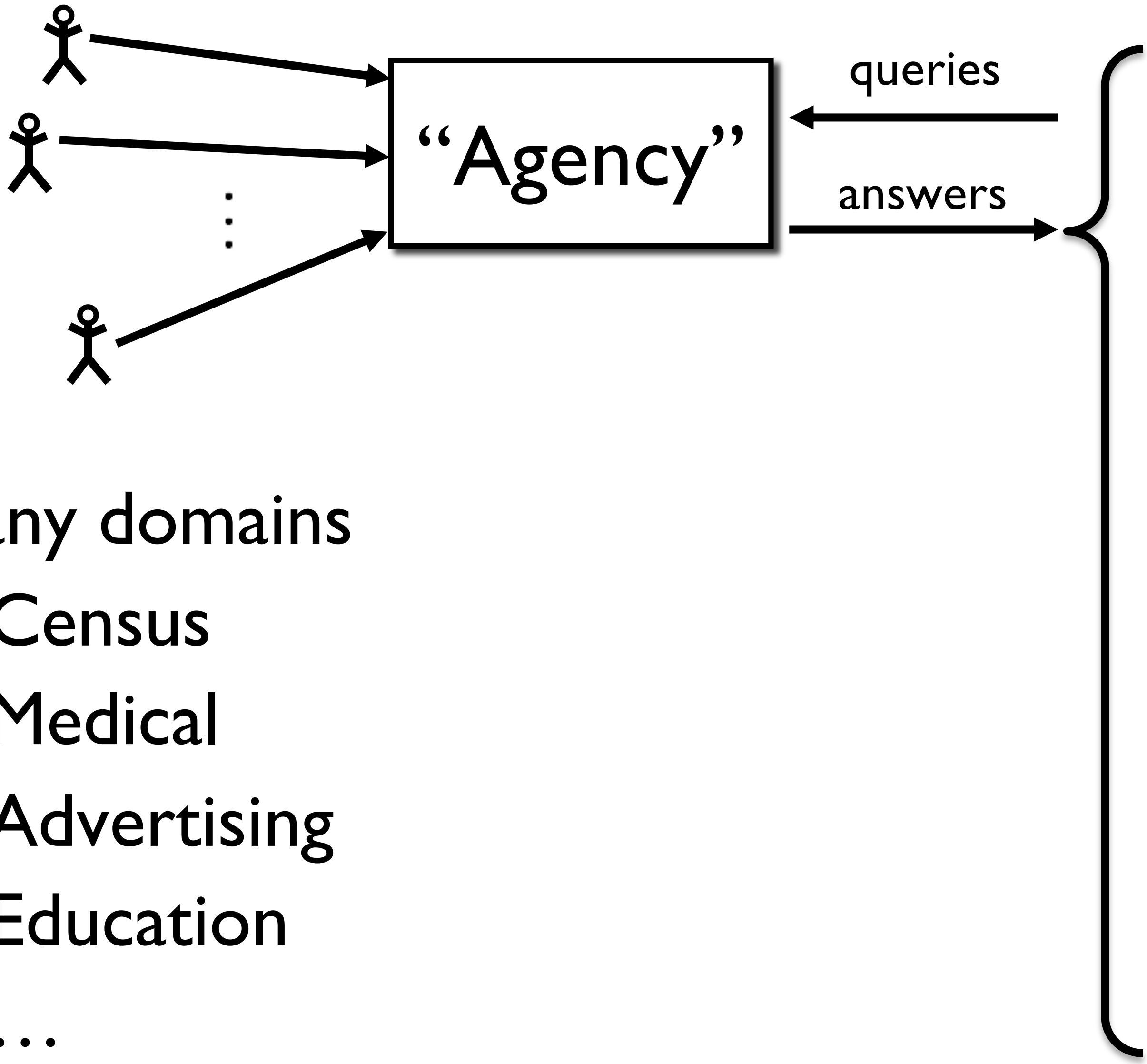
BOSTON UNIVERSITY

Based on
- L. Reyzin, A. Smith, S. Yakoubov
  https://eprint.iacr.org/2018/997
- A. Cheu, A. Smith, J. Ullman, D. Zeber, M. Zhilayev
  https://arxiv.org/abs/1808.01394

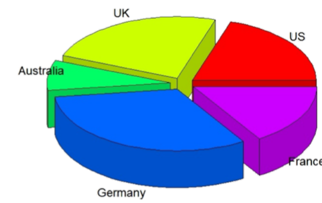
Privacy in Statistical Databases

Individuals

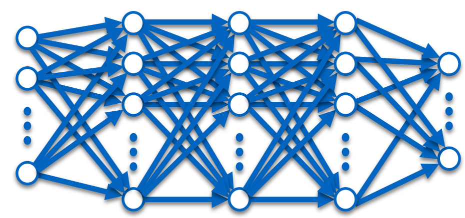
Researchers



Summaries



Complex models



Synthetic data

	Name	Birth Date	Country	State
1	Giovanni D'Agostini	3-Mar-1875	Italy	L'Aquila
2	Giovanni D'Agostini	3-Mar-1875	Italy	L'Aquila
3	Giovanni D'Agostini	3-Mar-1875	Italy	L'Aquila
4	Giovanni D'Agostini	3-Mar-1875	Italy	L'Aquila
5	Giovanni D'Agostini	3-Mar-1875	Italy	L'Aquila
6	Giovanni D'Agostini	3-Mar-1875	Italy	L'Aquila
7	Bill Smith	4-Apr-1956	United States	Texas
8	Bill Smith	4-Apr-1956	United States	Texas
9	Bill Smith	4-Apr-1956	United States	Texas
10	Bill Smith	4-Apr-1956	United States	Texas

…

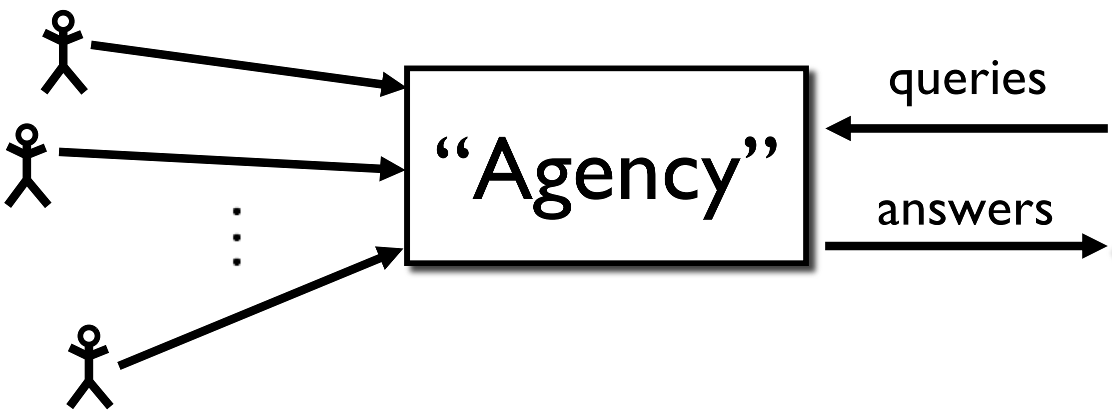
Many domains

- Census
- Medical
- Advertising
- Education
- …

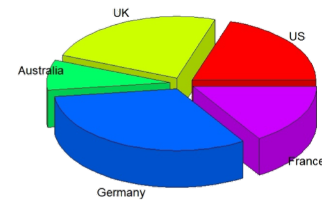
Privacy in Statistical Databases

Individuals

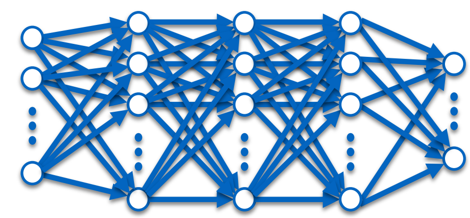
Researchers



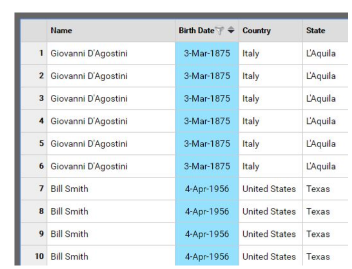
Summaries



Complex models



Synthetic data

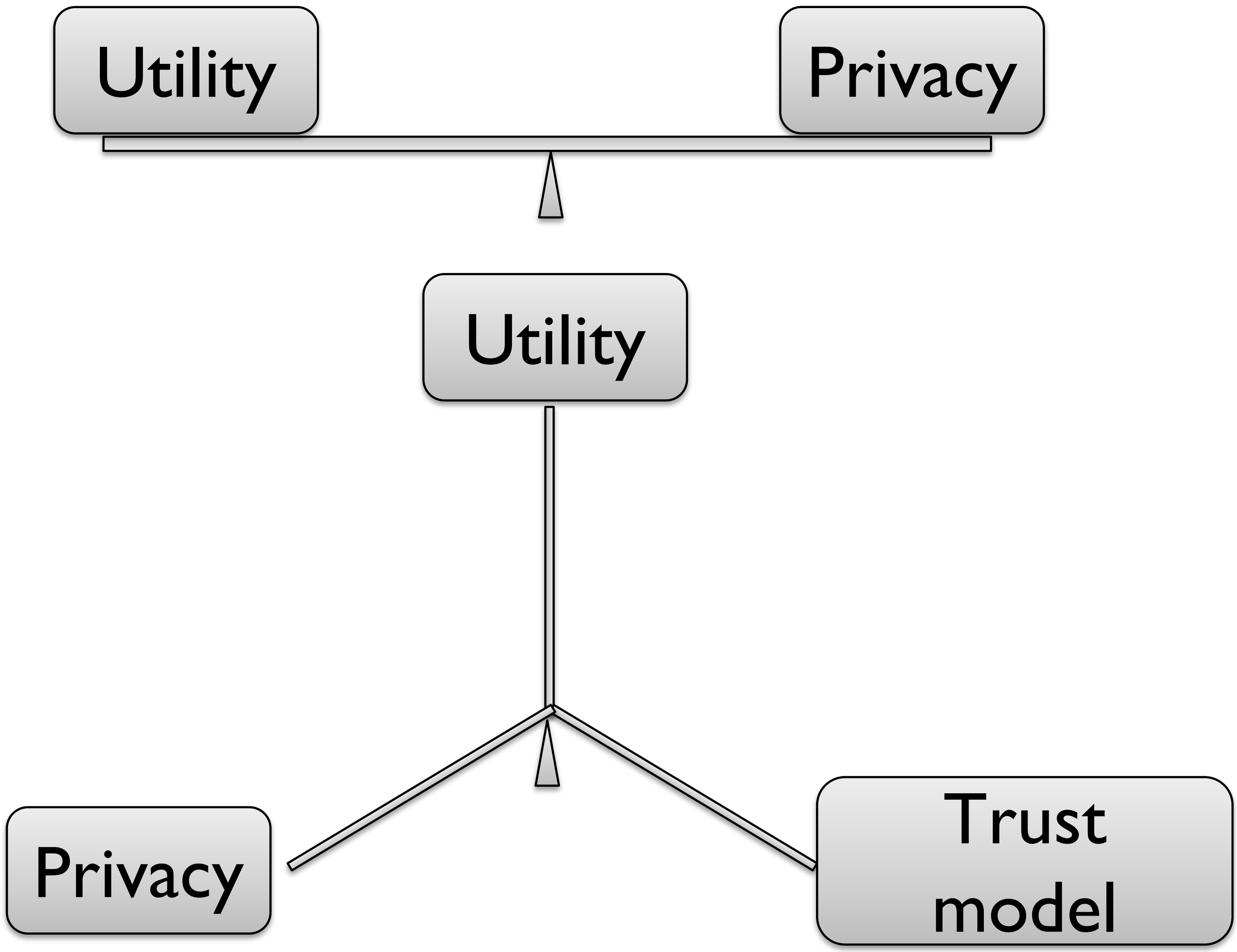


	Name	Birth Date	Country	State
1	Giovanni D'Agostini	3-Mar-1875	Italy	L'Aquila
2	Giovanni D'Agostini	3-Mar-1875	Italy	L'Aquila
3	Giovanni D'Agostini	3-Mar-1875	Italy	L'Aquila
4	Giovanni D'Agostini	3-Mar-1875	Italy	L'Aquila
5	Giovanni D'Agostini	3-Mar-1875	Italy	L'Aquila
6	Giovanni D'Agostini	3-Mar-1875	Italy	L'Aquila
7	Bill Smith	4-Apr-1956	United States	Texas
8	Bill Smith	4-Apr-1956	United States	Texas
9	Bill Smith	4-Apr-1956	United States	Texas
10	Bill Smith	4-Apr-1956	United States	Texas

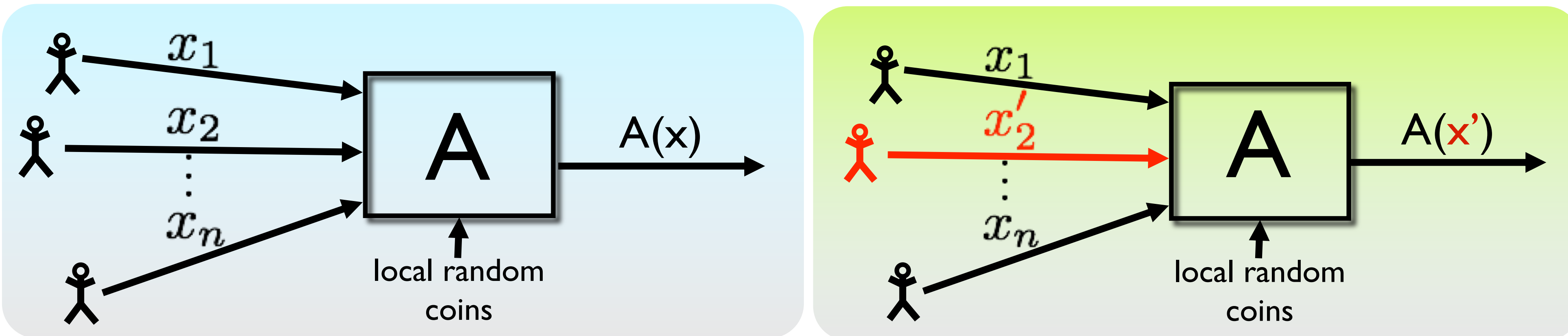
…

"Aggregate" outputs can leak lots of information

- Reconstruction attacks
- Example: Ian Goldberg's talk on "the secret sharer"


Utility
Privacy
Utility
Privacy
Trust model

Differential Privacy *[Dwork, McSherry, Nissim, S. 2006]*



x' is a neighbor of ***x***
if they differ in one data point

Neighboring databases induce **close** distributions on outputs

Definition: A is (ϵ, δ)-differentially private if
for all neighbors $\boldsymbol{x}$, $\boldsymbol{x}'$,
for all sets of outputs T

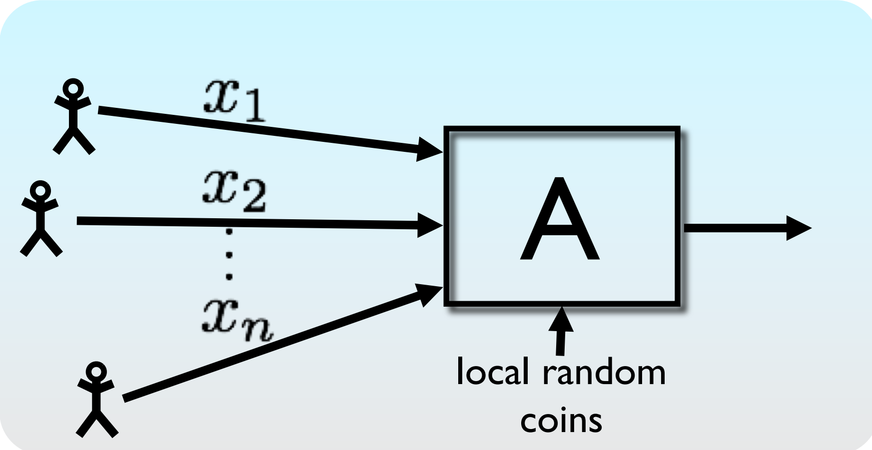
$$\Pr_{\text{coins of } A}(A(\boldsymbol{x}) \in T) \leq e^{\epsilon} \cdot \Pr_{\text{coins of } A}(A(\boldsymbol{x}') \in T) + \delta$$

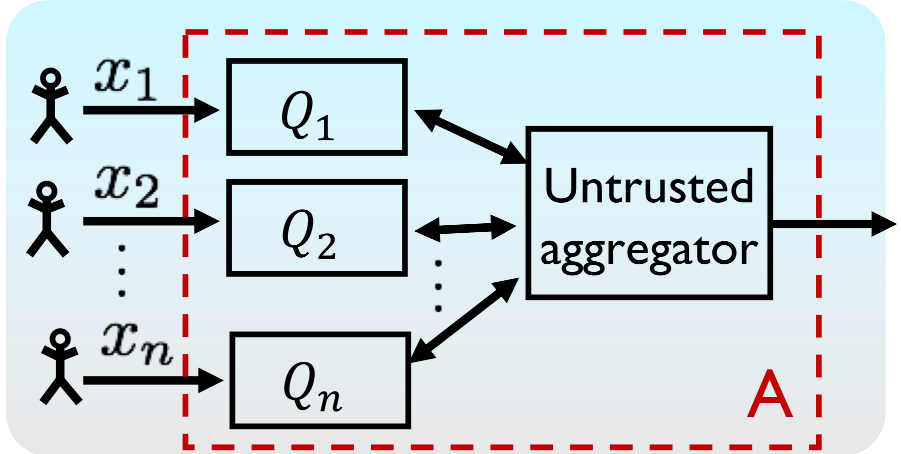
Outline

- Local model
- Models for DP + MPC
- Lightweight architectures
  - “From HATE to LOVE MPC”
- Minimal primitives
  - “Differential Privacy via Shuffling”

Local Model for Privacy

Equivalent to [Efvimievski, Gehrke, Srikant '03]





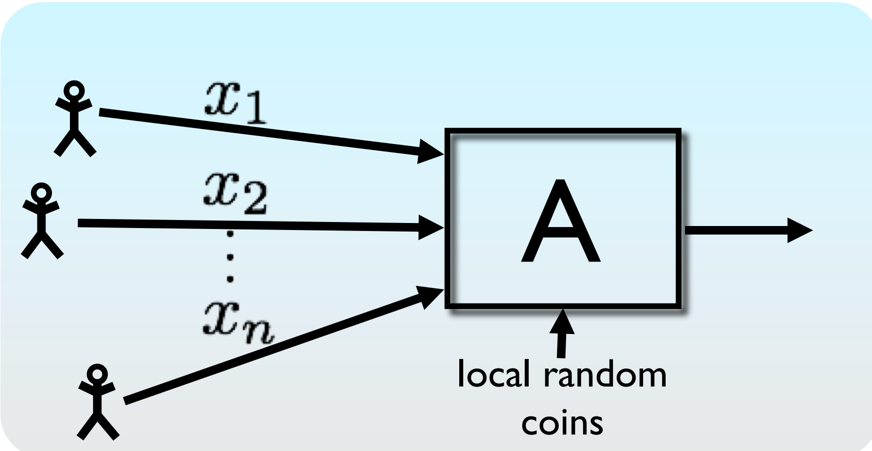
- "Local" model
  - Person i randomizes their own data
  - Attacker sees everything except player i's local state

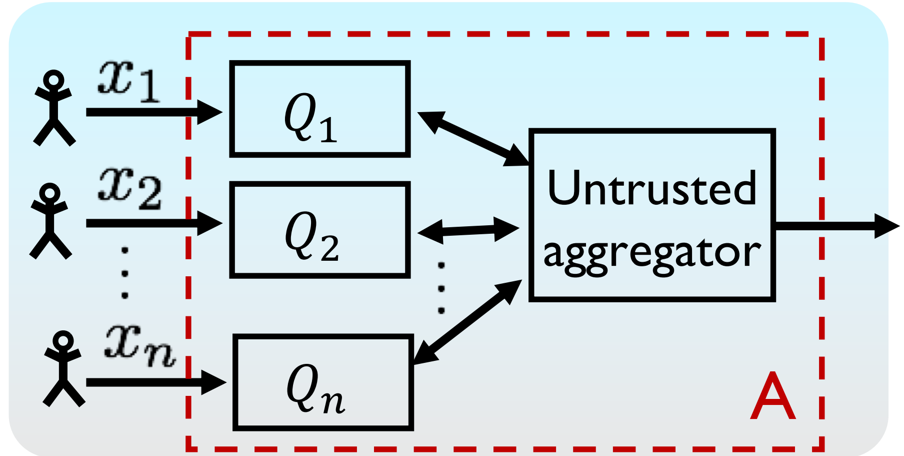
- Definition: A is ϵ-locally differentially private if for all i:
  - for all neighbors $\boldsymbol{x}$, $\boldsymbol{x}'$,
  - for all behavior B of other parties,
  - for all sets of transcripts T:

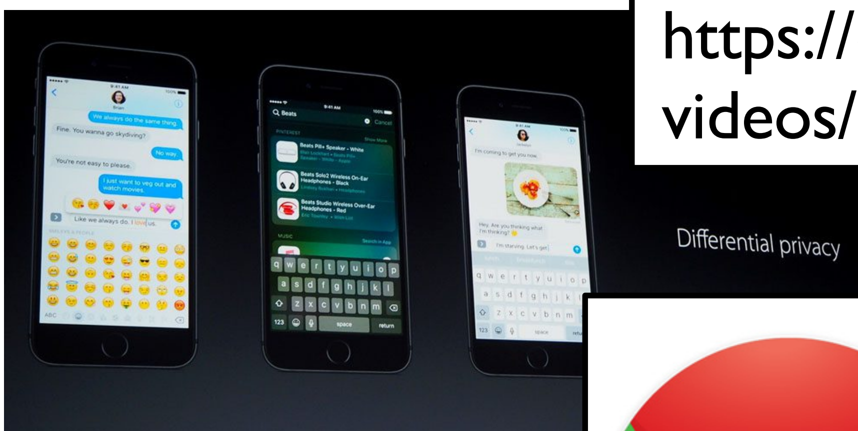
$$\Pr_{\text{coins } r_i}(A(\boldsymbol{x}, B) = t) \le e^{\epsilon} \cdot \Pr_{\text{coins } r_i}(A(\boldsymbol{x}', B) = t)$$

$\delta = 0$ w.l.o.g.

Local Model for Privacy



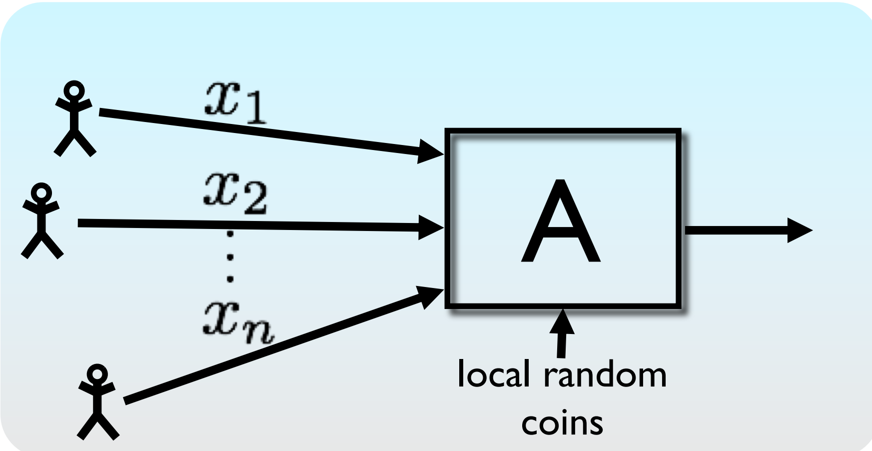


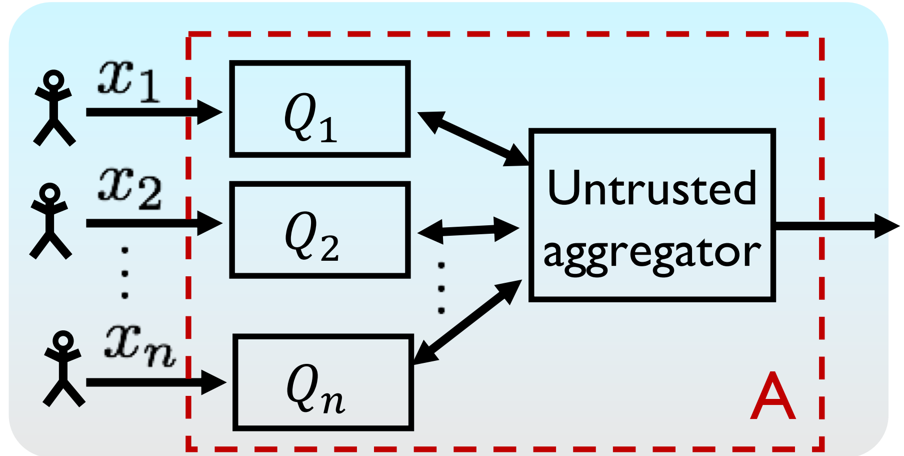


https://developer.apple.com/
videos/play/wwdc2016/709/

https://github.com/google/rappor

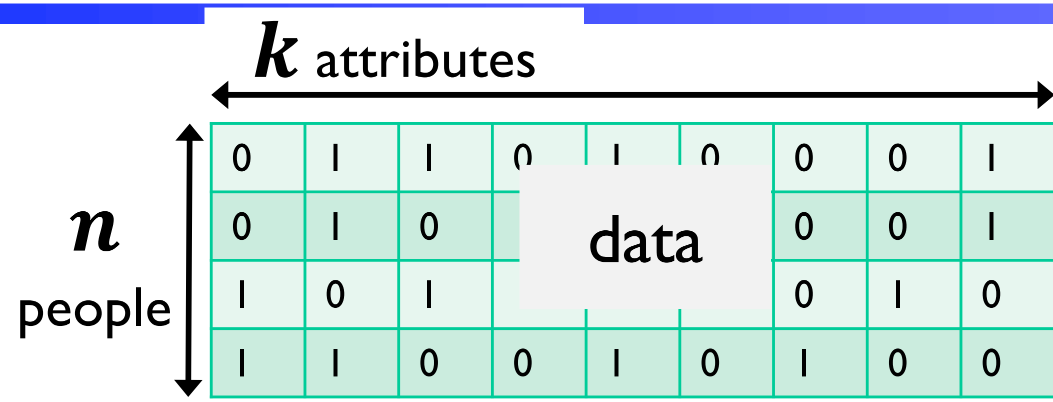
Local Model for Privacy





- Pros
  - No trusted curator
  - No single point of failure
  - Highly distributed
  - Beautiful algorithms
- Cons
  - Low accuracy
    - Proportions: $\Theta\left(\frac{1}{\epsilon\sqrt{n}}\right)$ error [BMO'08,CSS'12] vs $O(\frac{1}{n\epsilon})$ central
  - Correctness requires honesty

Selection Lower Bounds [DJW'13, Ullman '17]



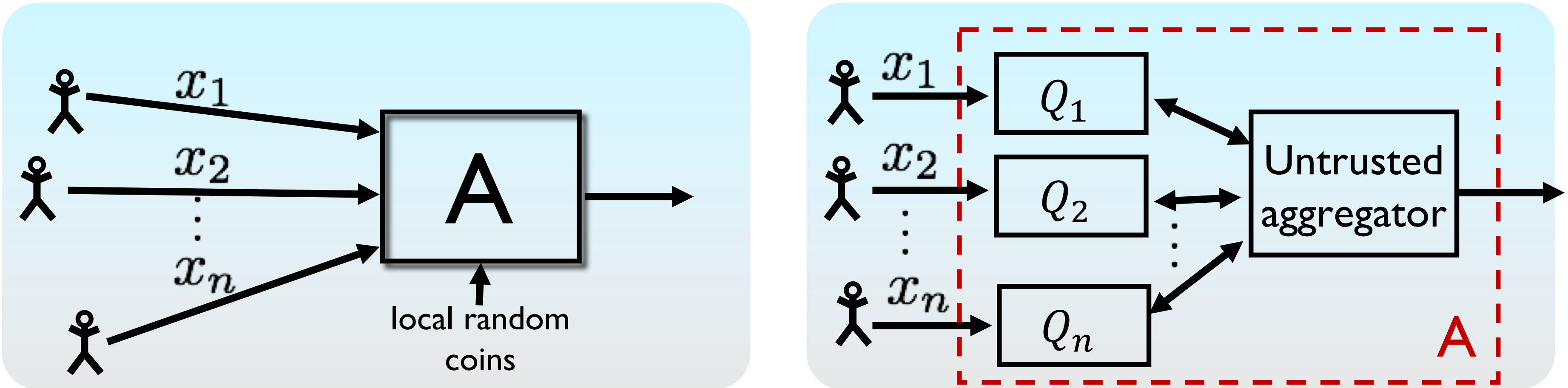
- Suppose each person has k binary attributes
- **Goal**: Find index j with highest count $(\pm\alpha)$

- **Central model**: $n = O(\log(k)/\epsilon\alpha)$ suffices
  [McSherry Talwar '07]
- **Local model:** Any **noninteractive** local DP protocol with nontrivial error requires

$$n = \Omega(k \log(k) / \epsilon^2)$$

  - [DJW'13, Ullman '17]
  - (No lower bound known for interactive protocols)

Local Model for Privacy

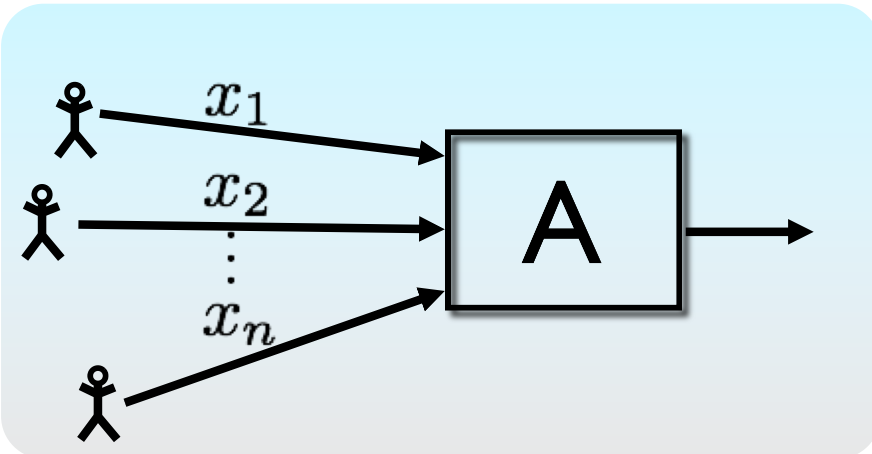


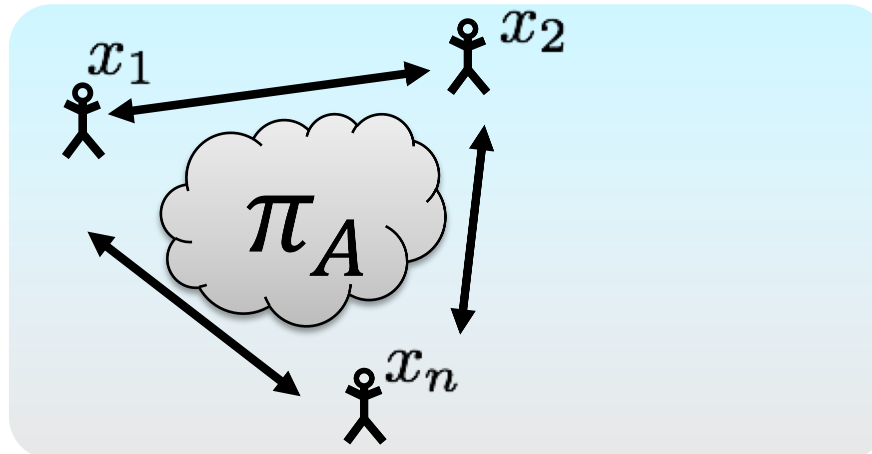
What other models allow similarly distributed trust?

Outline

- Local model
- Models for DP + MPC
- Lightweight architectures
  - “From HATE to LOVE MPC”
- Minimal primitives
  - “Differential Privacy via Shuffling”

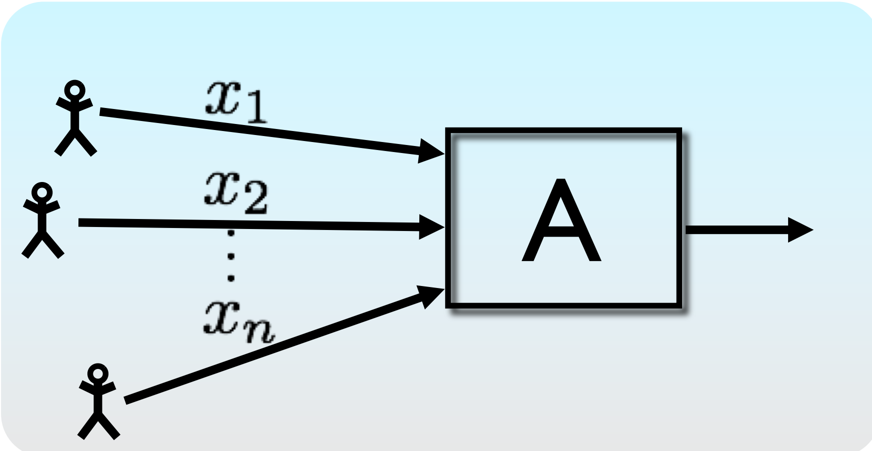
Two great tastes that go great together

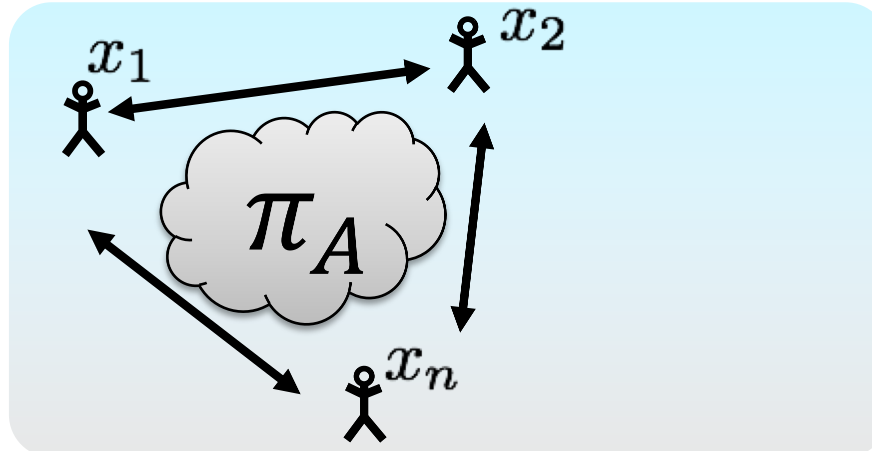




- How can we get accuracy without a trusted curator?
- Idea: Replace central algorithm A with multiparty computation (MPC) protocol for A (randomized), and either
  - Secure channels + honest majority
  - Computational assumptions + PKI
- Questions:
  - What definition does this achieve?
  - Are there special-purpose protocols that are more efficient than generic reductions?
  - What models make sense?
  - What primitives are needed?

Definitions





What definitions are achieved?

- Simulation of an (ϵ, δ)-DP protocol
- Computational DP [Mironov, Pandey, Reingold, Vadhan'08]

Not equivalent

Definition: A is (t, ϵ, δ)-computationally differentially private if,
for all neighbors x, x',
for all distinguishers $T \in time(t)$

$$\Pr_{\text{coins of } A}\big(T\big(A(x)\big) = 1\big) \leq e^{\epsilon} \cdot \Pr_{\text{coins of } A}\big(T\big(A(x')\big) = 1\big) + \delta$$

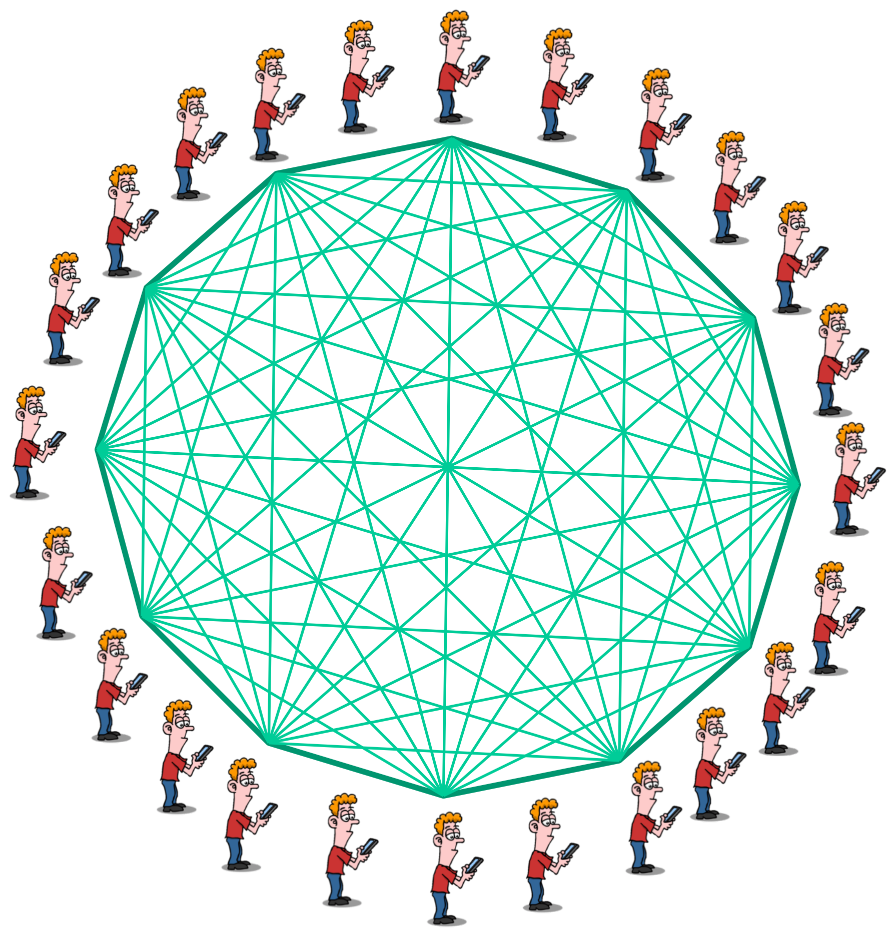
Question 1: Special-purpose protocols

- [Dwork Kenthapadi McSherry Mironov Naor ‘06]
  Special-purpose protocols for generating Laplace/exponential noise via finite field arithmetic
  - ⇒ honest-majority MPC
  - Satisfies simulation, follows existing MPC models
  - Lots of follow-up work

- [He, Machanavajjhala, Flynn, Srivastava ’17, Mazloom, Gordon ’17, maybe others?]
  Use DP statistics to speed up MPC
  - Leaks more than ideal functionality

Question 2: What MPC models make sense?

- Recall: secure MPC protocols require
  - Communication between all pairs of parties
  - Multiple rounds, so parties have to stay online

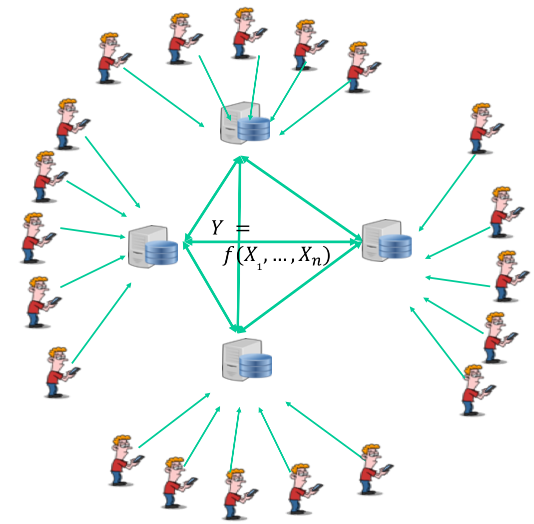


- Protocols involving all Google/Apple users wouldn't work

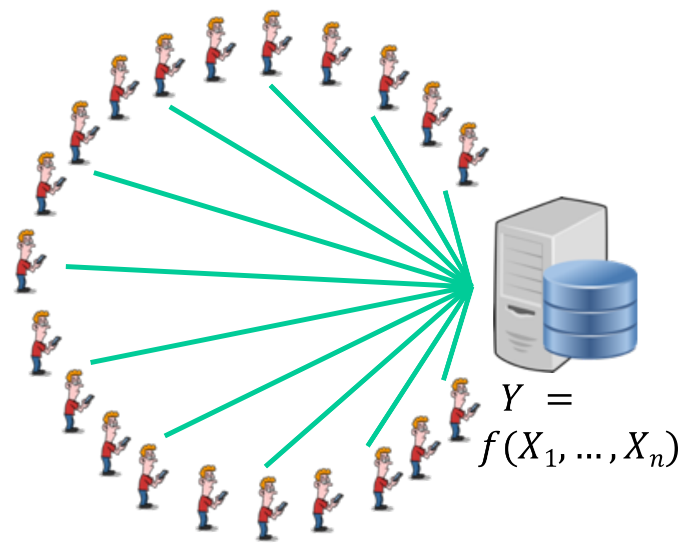
Question 2: What MPC models make sense?

Applications of DP suggest a few different settings

- "Few hospitals"
  - ➢ Small set of computationally powerful data holders
  - ➢ Each holds many participants' data
  - ➢ Data holders have their own privacy-related concerns
    - Sometimes can be modeled explicitly, e.g. [Haney, Machanavajjhala, Abowd, Graham, Kutzbach, Vilhuber '17]
    - Data holders interests may not align with individuals'



- "Many phones"
  - ➢ Many weak clients (individual data holders)
  - ➢ One server or small set of servers
  - ➢ Unreliable, client-server network
  - ➢ Calls for lightweight MPC protocols, e.g. [Shi, Chan, Rieffel, Chow, Song '11, Boneh, Corrigan-Gibbs '17, Bonawitz, Ivanov, Kreuter, Marcedone, McMahan, Patel, Ramage, Segal, Seth '17]

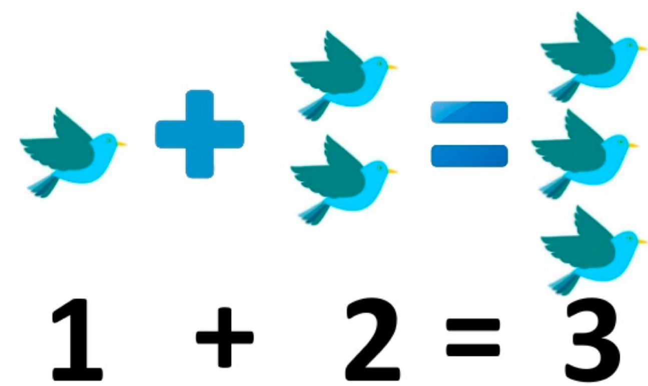


DP does not need full MPC

- ➢ Sometimes, leakage helps [HMFS '17, MG'17]
- ➢ Sometimes, we do not know how to take advantage of it [McGregor Mironov Pitassi Reingold Talwar Vadhan '10]

Question 3: What MPC primitives do we need?

- Observation: Most DP algorithms rely on 2 primitives
  - **Addition** + Laplace/Gaussian noise
  - **Threshold**(summation + noise)
    - Sufficient for "sparse vector" and "exponential mechanism"
- [Shafi's talk mentions others for training nonprivate deep nets.]
  - Relevant for PATE framework



- Lots of work focuses on addition
  - "Federated learning"
  - Relies on users to introduce small amounts of noise
- Thresholding remains complicated
  - Because highly nonlinear
  - Though maybe approximate thresholding easier (e.g. HEEAN)
- Recent papers look at weaker primitives
  - Shufflers as a useful primitive [Erlingsson, Feldman, Mironov, Raghunathan, Talwar, Thakurta] [Cheu, Smith, Ullman, Zeber, Zhilyaev 2018]

Outline

- Local model
- Models for DP + MPC
- Lightweight architectures
  - “From HATE to LOVE MPC”
- Minimal primitives
  - “Differential Privacy via Shuffling”

Turning HATE into LOVE MPC

Scalable Multi-Party Computation With Limited Connectivity

Leonid Reyzin, Adam Smith, Sophia Yakoubov

https://eprint.iacr.org/2018/997

Goals

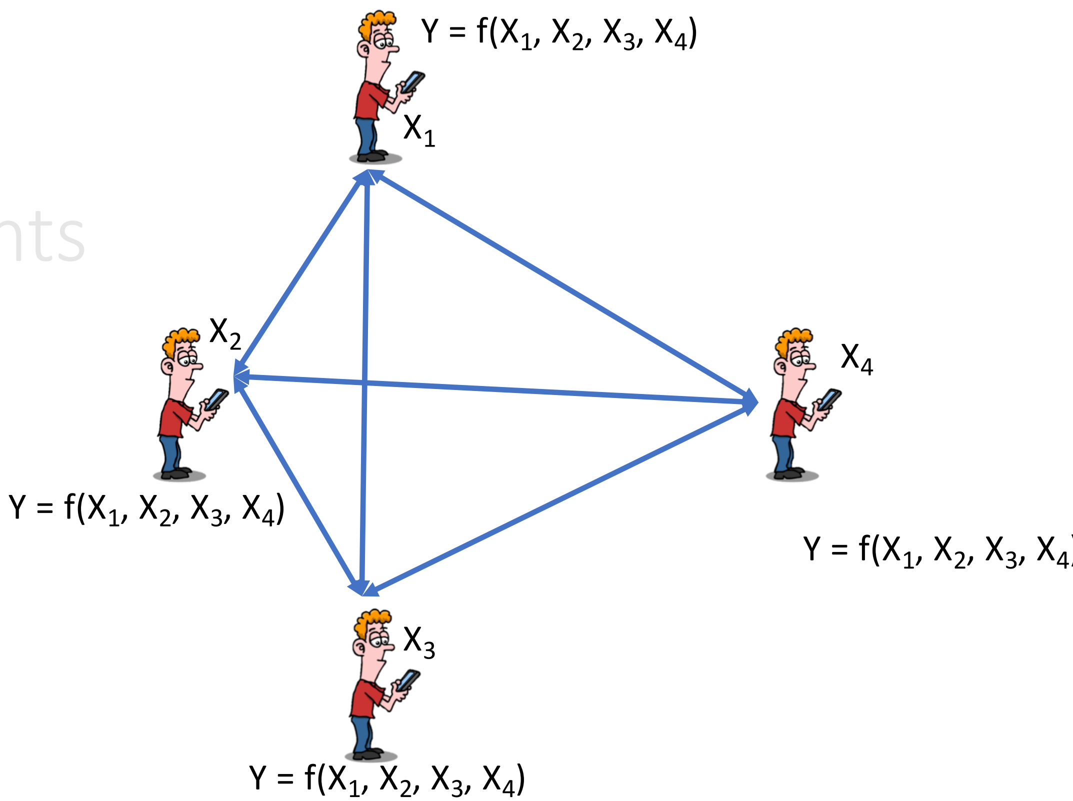
- Clean formalism for “many phones” model
  - Inspired by protocols of [Shi et al, 2011; Bonawitz et al. 2017]

- Identify
  - Fundamental limits
  - Potentially practical protocols
  - Open questions

Large-scale
One-server
Vanishing-participants
Efficient

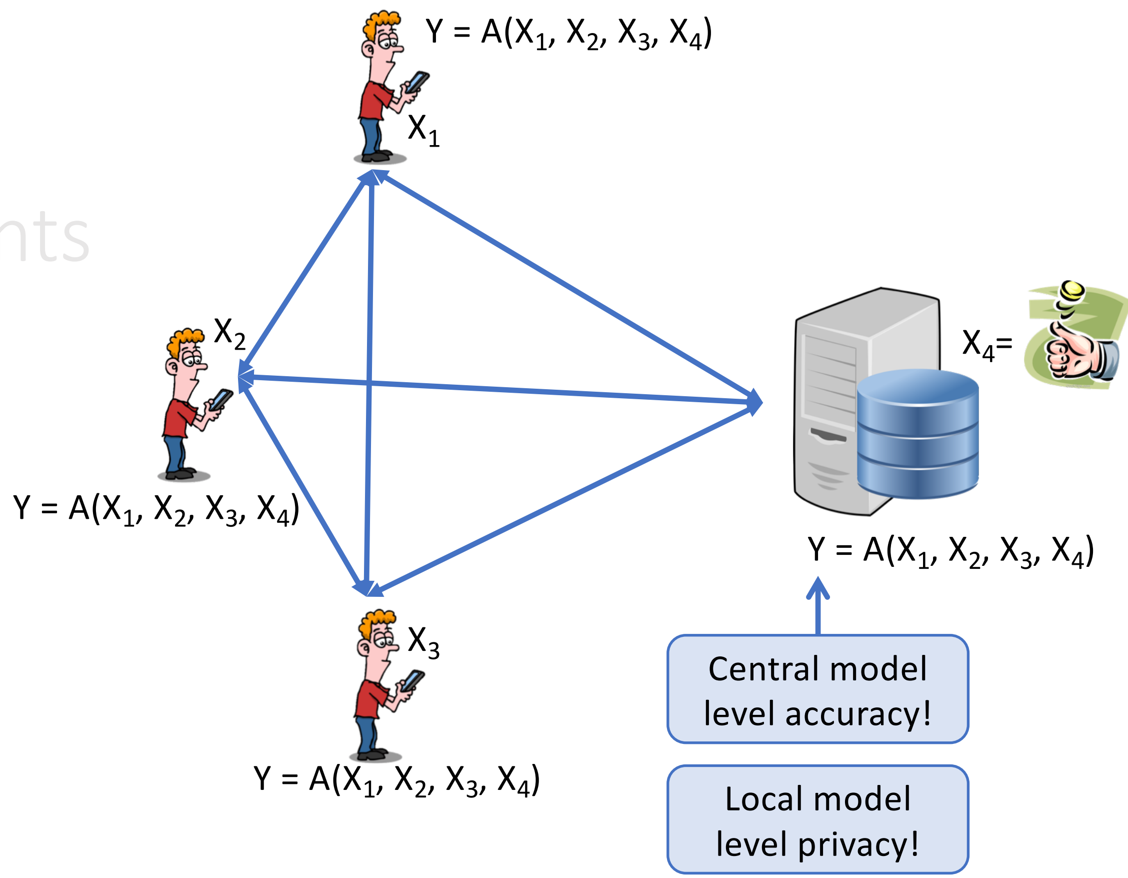
MPC

[Goldreich,Micali,Widgerson87,Yao87]



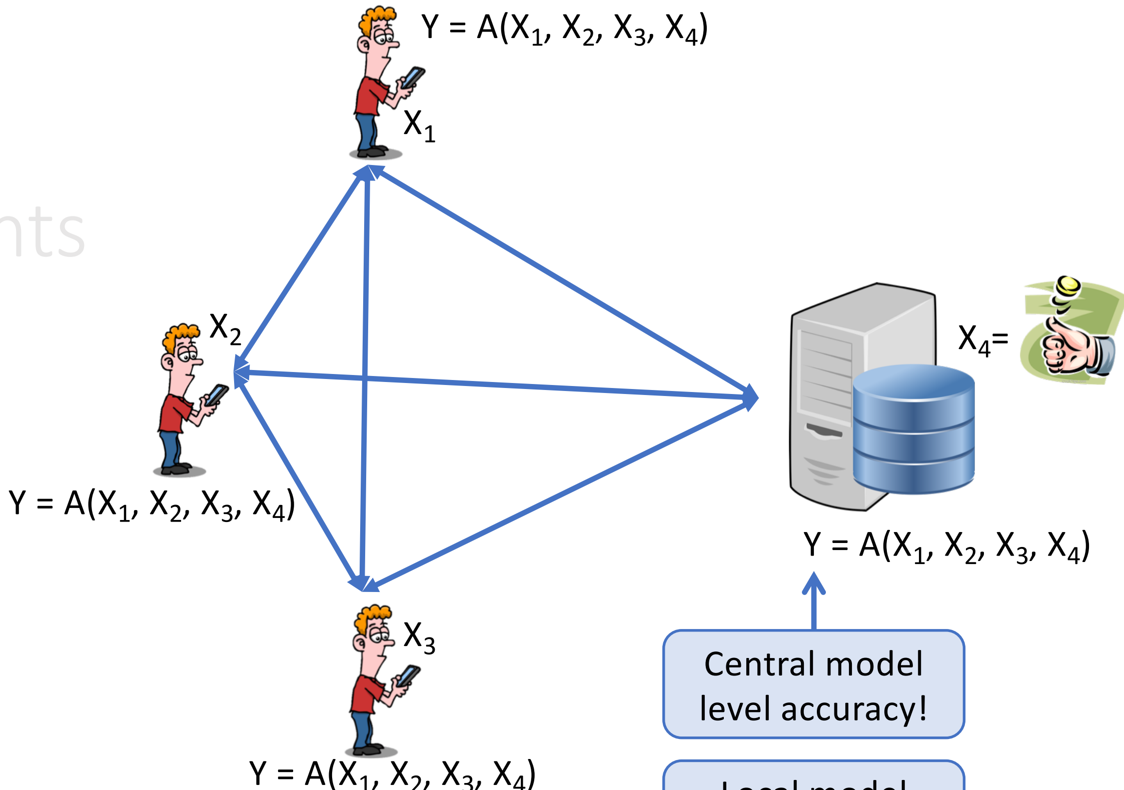
No party learns anything other than the output!

Large-scale
One-server
Vanishing-participants
Efficient
MPC



Can compute differentially private statistic A(X) without server learning anything but the output!
[Dwork,Kenthapadi,McSherry,Mironov,Naor06]

Large-scale
One-server
Vanishing-participants
Efficient
MPC

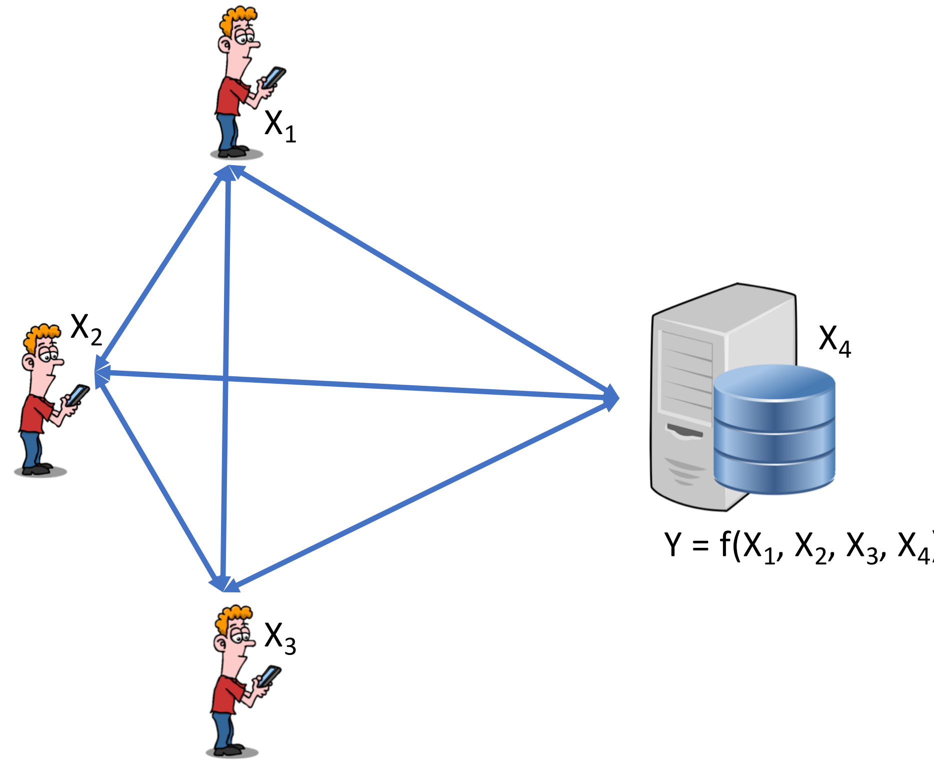


Central model level accuracy!

Local model level privacy!

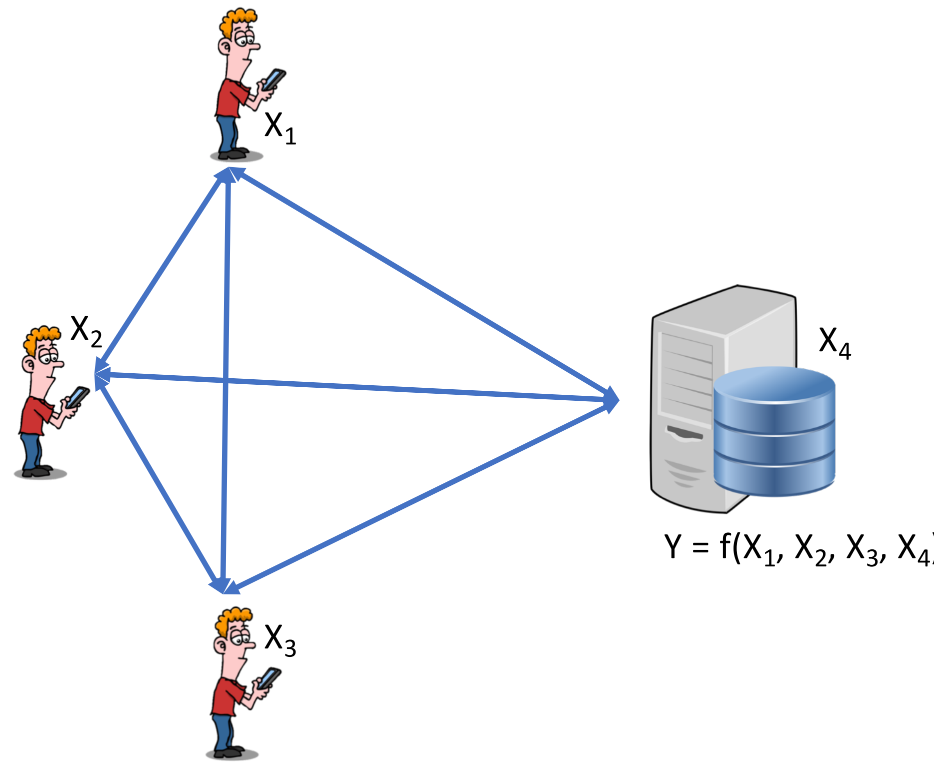
Can compute differentially private statistic A(X) without server learning anything but the output!
A(X) is often linear, so we will focus on MPC for addition

Large-scale
One-server
Vanishing-participants
Efficient
MPC



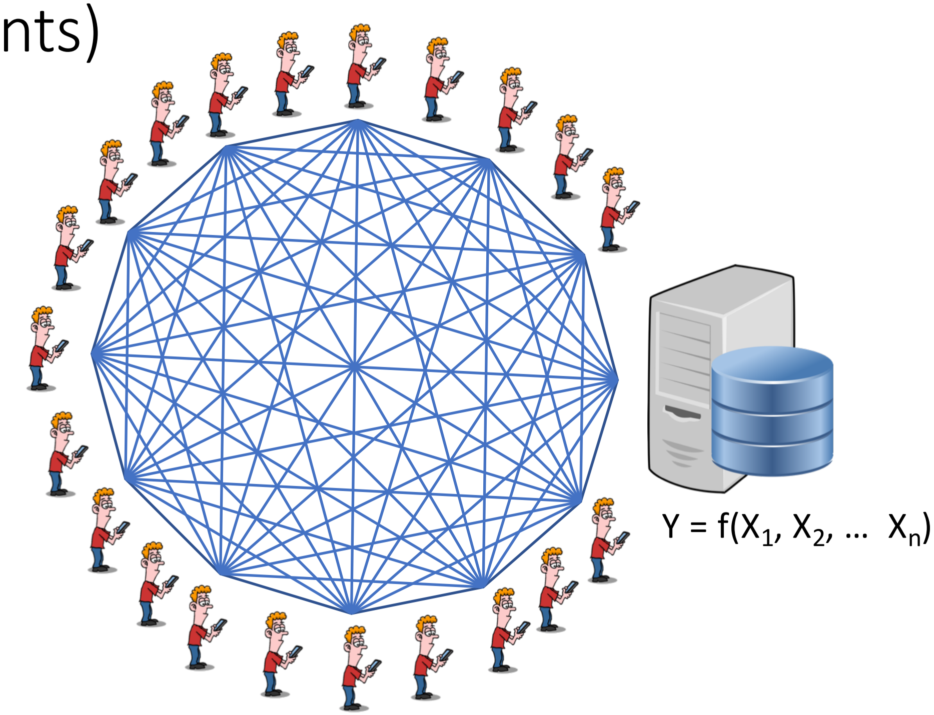
	Clients	Server
Computational power	weak	strong

Large-scale
One-server
Vanishing-participants
Efficient
MPC



	Clients	Server
Computational power	weak	strong

Large-scale (millions of clients)
One-server
Vanishing-participants
Efficient
MPC



	Clients	Server
Computational power	weak	strong

Large-scale (millions of clients)

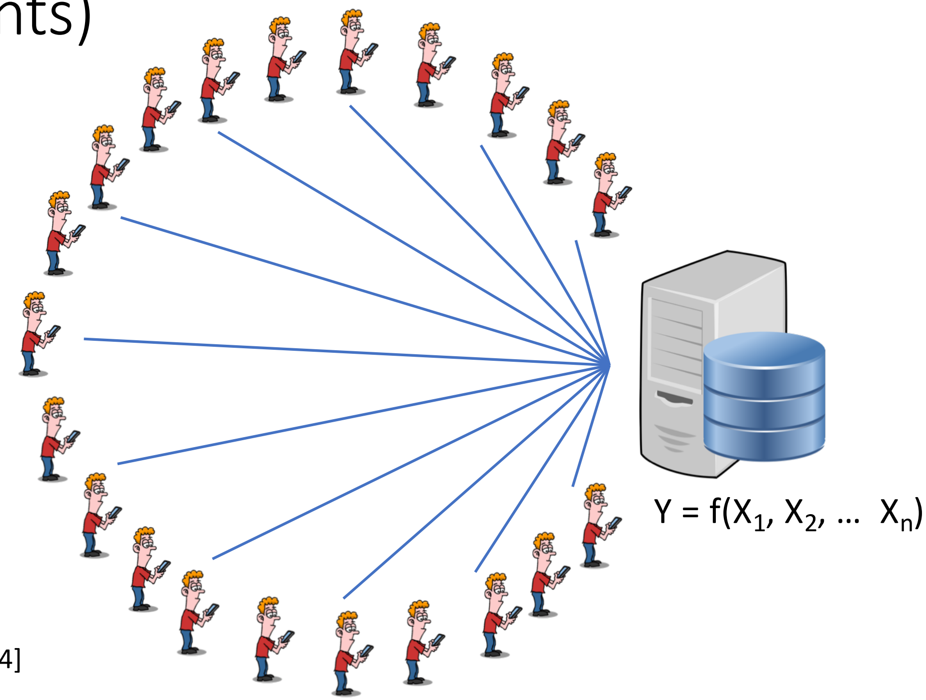
One-server

Vanishing-participants

Efficient

MPC

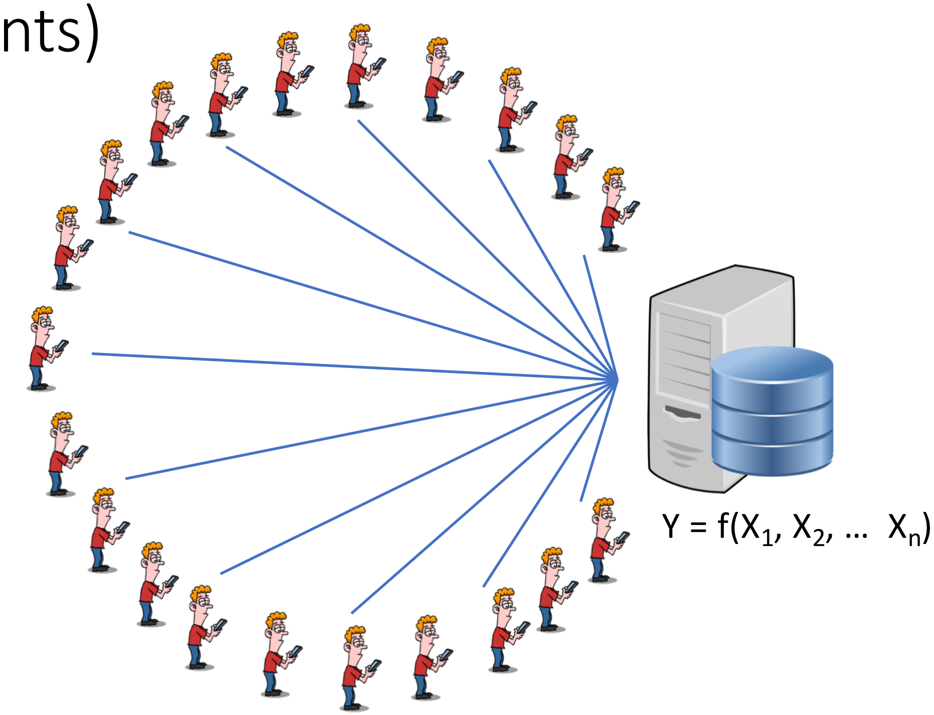
- Star communication graph, as in noninteractive multiparty computation (NIMPC) [Beimel,Gabizon,Ishai,Kushilevitz,Meldgaard,PaskinCherniavsky14]



	Clients	Server
Computational power	weak	strong
Direct communication	only to server	to everyone

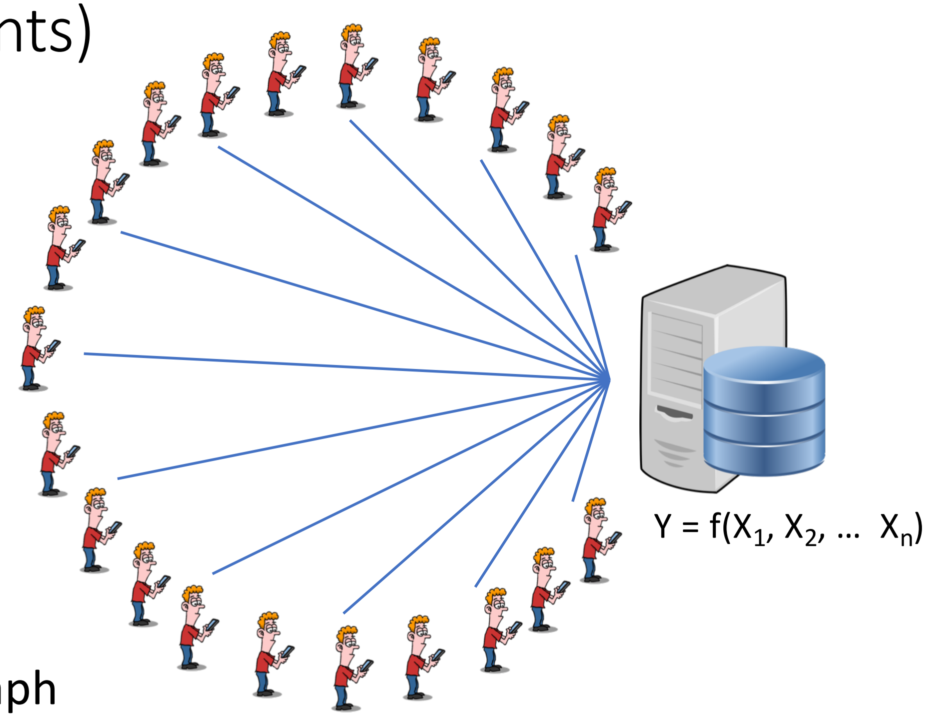
Large-scale (millions of clients)
One-server
Vanishing-participants
Efficient
MPC

- Computation must complete even if some clients abort



	Clients	Server
Computational power	weak	strong
Direct communication	only to server	to everyone
Network	unreliable	reliable

Large-scale (millions of clients)
One-server
Vanishing-participants
Efficient
MPC



- Computation must complete even if some clients abort
  - Considered in many papers in the all-to-all communication graph [Badrinarayanan,Jain,Manohar,Sahai18]
  - Considered in [Bonawitz,Ivanov,Kreuter,Mercedone,McMahan,Patel,Ramage,Segal,Seth17] in star communication graph, achieved in 5 message flows

What's the best we can do?

Our Contributions

- Defining LOVE MPC
- Minimal requirements for LOVE MPC:
  - 3 flows
  - Setup: correlated randomness of PKI
- Building LOVE MPC for addition
  - Main Tool: **H**omomorphic **A**d hoc **T**hreshold **E**ncryption
- Tradeoffs in LOVE MPC

Our Contributions

- Defining LOVE MPC
- Minimal requirements for LOVE MPC:
  - 3 flows
  - Setup: correlated randomness or **PKI**
- Building LOVE MPC for addition
  - Main Tool: **H**omomorphic **A**d hoc **T**hreshold **E**ncryption
- Tradeoffs in LOVE MPC

Our Contributions

- Defining LOVE MPC
- Minimal requirements for LOVE MPC:
  - 3 flows
  - Some setup: PKI
- Building LOVE MPC for addition
  - Main Tool: **H**omomorphic **A**d hoc **T**hreshold **E**ncryption
    - Definitions
    - Construction: Share-And-Encrypt
  - Putting it all together
- Tradeoffs in LOVE MPC

LOVE MPC for Addition

	PK Size	Communication Per Party	Message Space Size	Assumption Family
[BIKMMPRSS17]	O(1)	O(n)	any	

LOVE MPC for Addition

	PK Size	Communication Per Party	Message Space Size	Assumption Family
[BIKMMPRSS17]	O(1)	O(n)	any	
Fully Homomorphic ATE [Badrinarayanan, Jain, Manohar, Sahai 2018]	O(1)	poly(n)	any	lattices
Shamir-and-ElGamal	O(1)	O(n)	small	DDH
CRT-and-Paillier	O(1)	O(n)	any	factoring
Obfuscation	poly(n)	O(1)	small	iO

LOVE MPC from HATE

OUR WORK

LOVE MPC for Addition

LOVE MPC from HATE

OUR WORK

	PK Size	Communication Per Party	Message Space Size	Assumption Family	Number of Rounds
[BIKMMPRSS17]	O(1)	O(n)	any		5
Fully Homomorphic ATE [Badrinarayanan, Jain, Manohar, Sahai 2018]	O(1)	poly(n)	any	lattices	3
Shamir-and-ElGamal	O(1)	O(n)	small	DDH	3
CRT-and-Paillier	O(1)	O(n)	any	factoring	3
Obfuscation	poly(n)	O(1)	small	iO	3

LOVE MPC for Addition

	PK Size	Communication Per Party		Message Space Size	Assumption Family	Number of Rounds	
		1st	nth			1st	nth
[BIKMMPRSS17]	O(1)	O(n)	O(n)	any		5	5
Fully Homomorphic ATE [Badrinarayanan, Jain, Manohar, Sahai 2018]	O(1)	poly(n)	poly(n)	any	lattices	3	3
Shamir-and-ElGamal	O(1)	O(n)	O(n)	small	DDH	3	3
CRT-and-Paillier	O(1)	O(n)	O(n)	any	factoring	3	3
Obfuscation	poly(n)	O(1)	O(1)	small	iO	3	3

LOVE MPC from HATE

OUR WORK

LOVE MPC for Addition

	PK Size	Communication Per Party		Message Space Size	Assumption Family	Number of Rounds	
		1st	nth			1st	nth
[BIKMMPRSS17]	O(1)	O(n)	O(n)	any		5	5
Fully Homomorphic ATE [Badrinarayanan, Jain, Manohar, Sahai 2018]	O(1)	poly(n)	poly(n)	any	lattices	3	3
Shamir-and-ElGamal	O(1)	O(n)	O(n)	small	DDH	3	3
CRT-and-Paillier	O(1)	O(n)	O(n)	any	factoring	3	3
Obfuscation	poly(n)	O(1)	O(1)	small	iO	3	3
Threshold ElGamal	O(1)	O(n)	O(1)	small	DDH	5	3

LOVE MPC from HATE

OUR WORK

Open Questions

	PK Size	Communication Per Party		Message Space Size	Assumption Family	Number of Rounds	
		1st	nth			1st	nth
[BIKMMPRSS17]	O(1)	O(n)	O(n)	any		5	5
Fully Homomorphic ATE [Badrinarayanan, Jain, Manohar, Sahai 2018]	O(1)	poly(n)	poly(n)	any	lattices	3	3
Shamir-and-ElGamal	O(1)	O(n)	O(n)	small	DDH	3	3
CRT-and-Paillier	O(1)	O(n)	O(n)	any	factoring	3	3
Obfuscation	poly(n)	O(1)	O(1)	small	iO	3	3
Threshold ElGamal	O(1)	O(n)	O(1)	small	DDH	5	3
?	O(1)	O(1)	O(1)			3	3

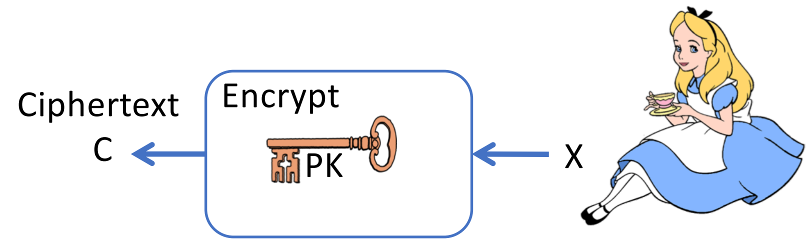
LOVE MPC from HATE

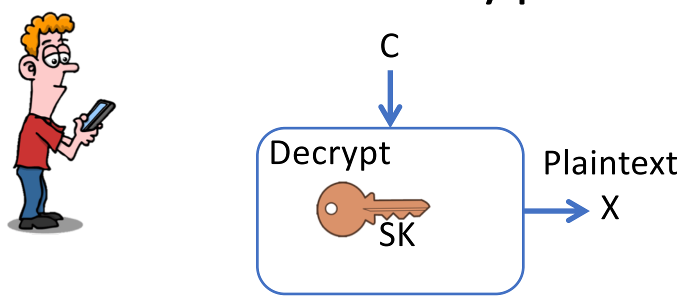
OUR WORK

Our Contributions

- Defining LOVE MPC
- Minimal requirements for LOVE MPC:
  - 3 flows
  - Some setup: PKI
- Building LOVE MPC for addition
  - Main Tool: **H**omomorphic **A**d hoc **T**hreshold **E**ncryption
    - Definitions
    - Construction: Share-And-Encrypt
  - Putting it all together
- Tradeoffs in LOVE MPC

Homomorphic
Ad Hoc
Threshold Encryption

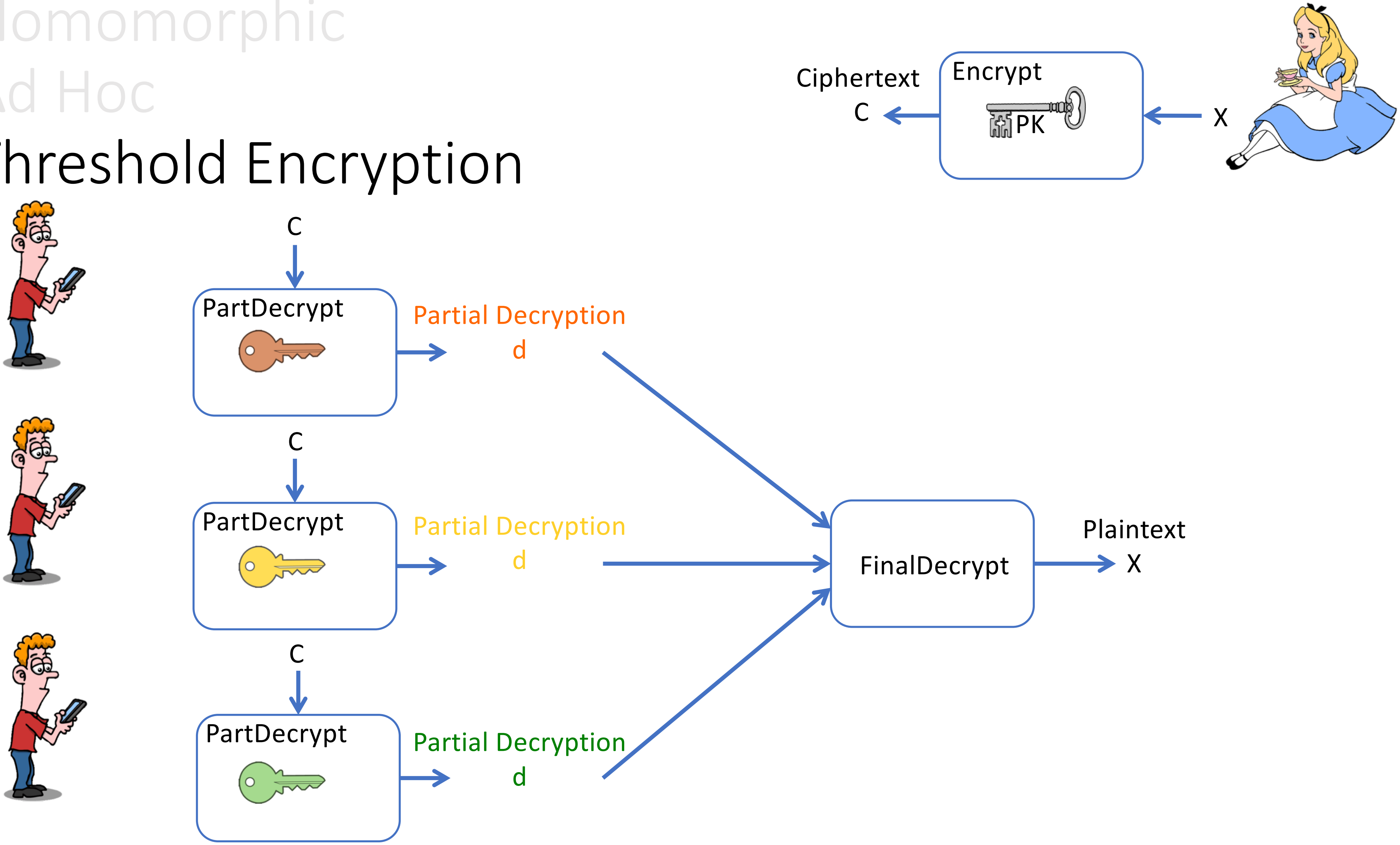




Homomorphic

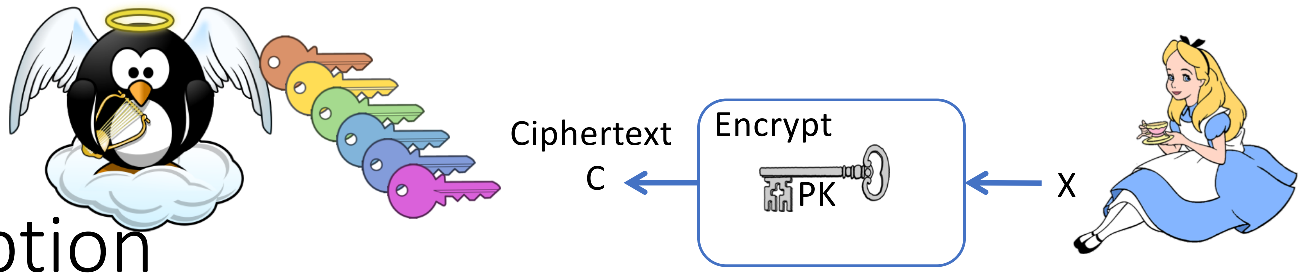
Ad Hoc

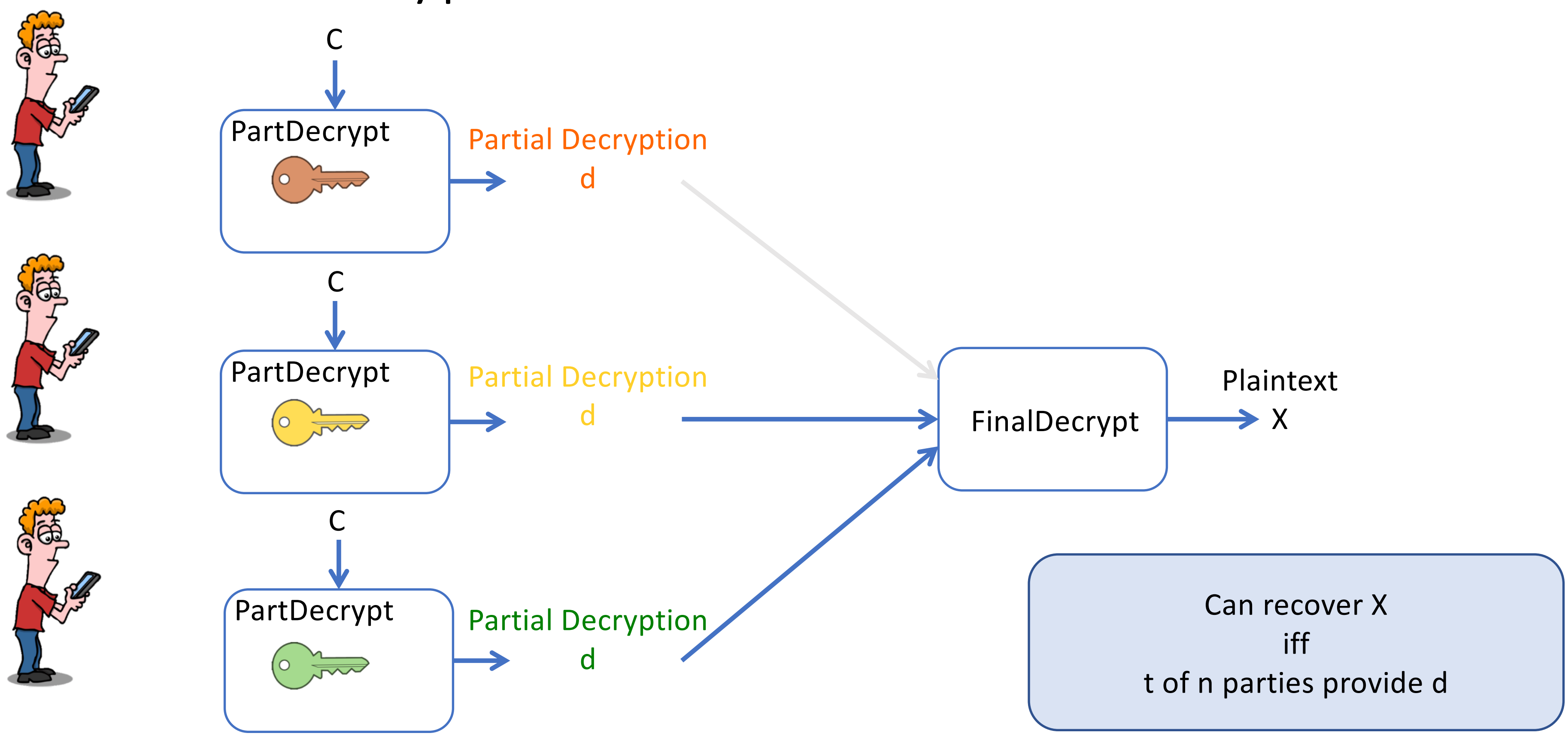
Threshold Encryption



Homomorphic
Ad Hoc

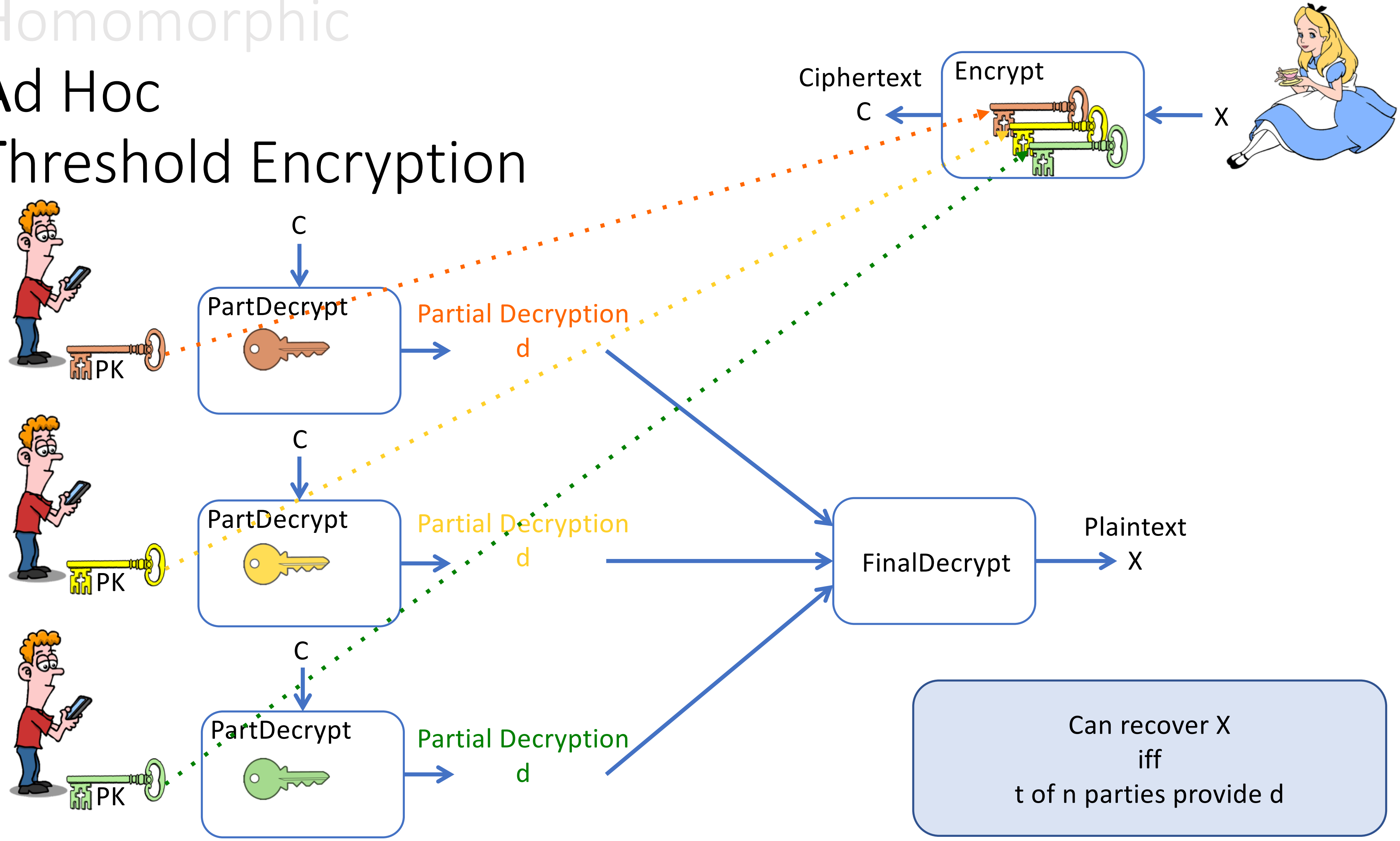
Threshold Encryption



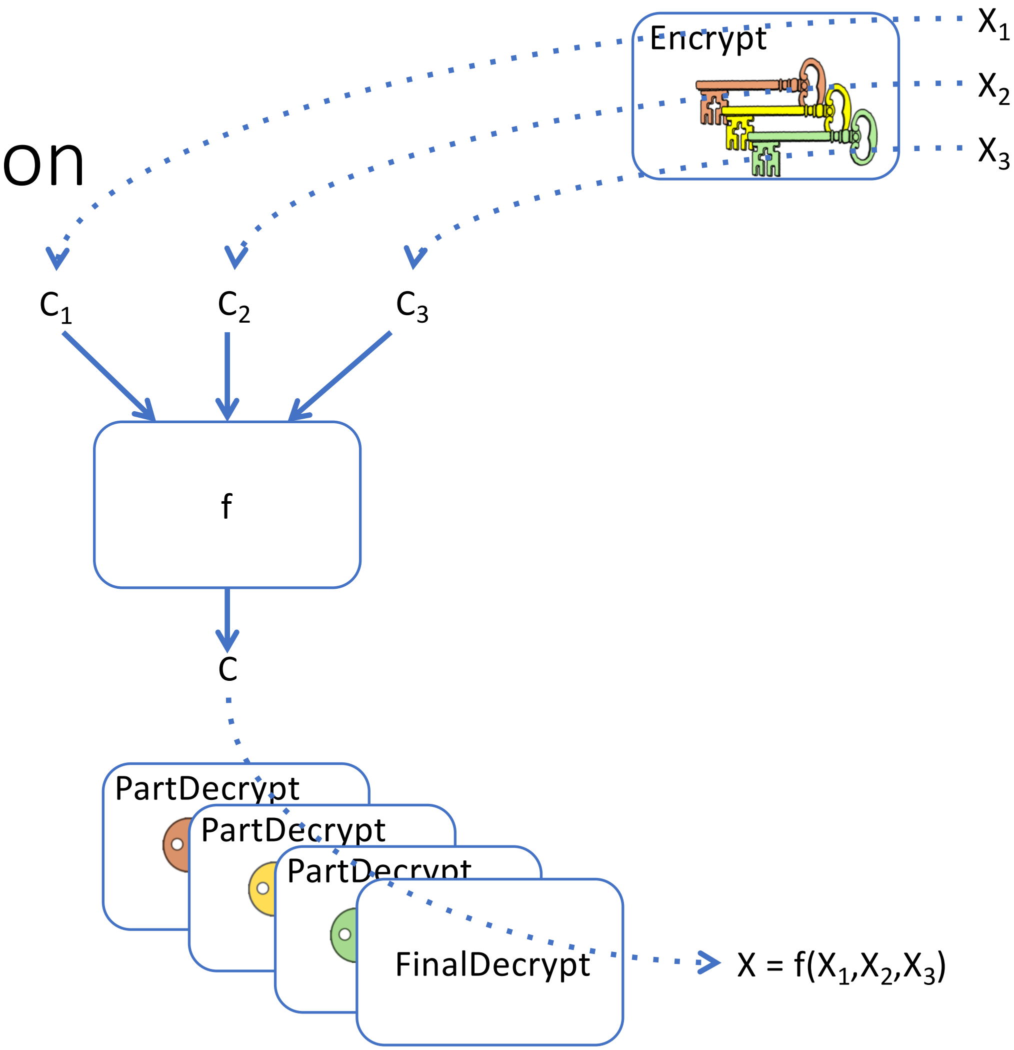


Homomorphic

Ad Hoc
Threshold Encryption



Homomorphic
Ad Hoc
Threshold Encryption



Additive HATE

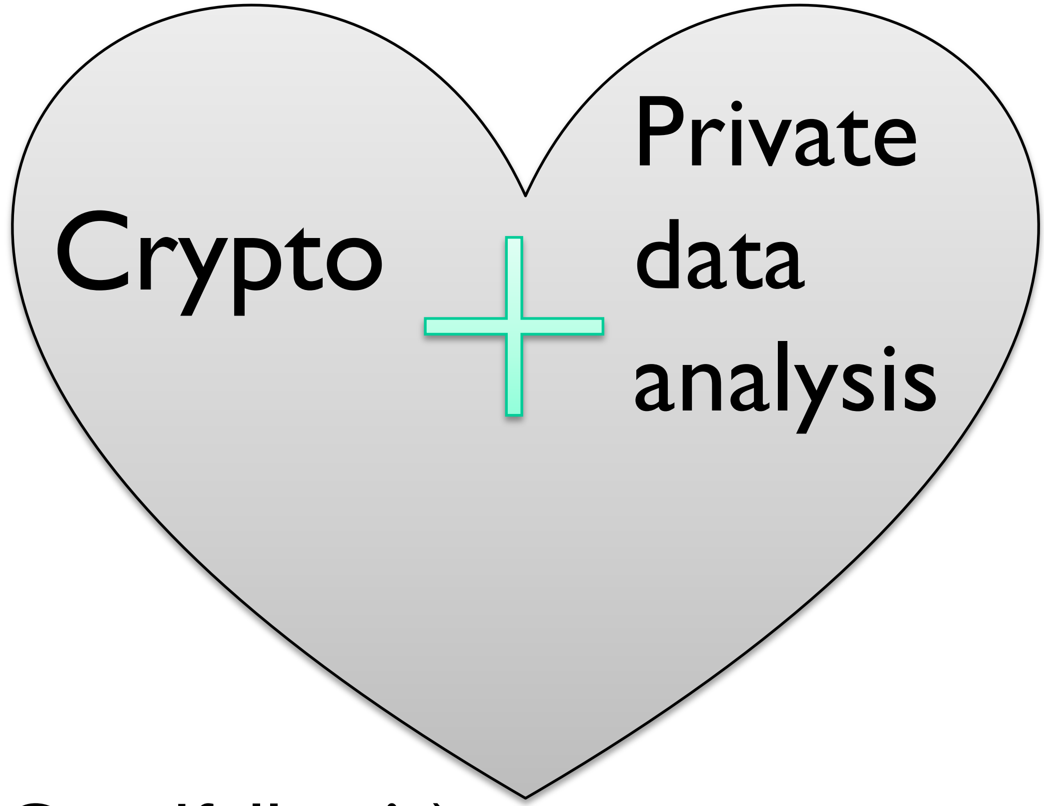
<table>
<tr><th></th><th></th><th>PK Size</th><th>Ciphertext Size</th><th>Message Space Size</th><th>Assumption Family</th></tr>
<tr><td></td><td>Fully Homomorphic ATE
[Badrinarayanan, Jain, Manohar, Sahai 2018]</td><td>O(1)</td><td>poly(n)</td><td>any</td><td>lattices</td></tr>
<tr><td rowspan="3">OUR WORK</td><td>Shamir-and-ElGamal</td><td>O(1)</td><td>O(n)</td><td>small</td><td>DDH</td></tr>
<tr><td>CRT-and-Paillier</td><td>O(1)</td><td>O(n)</td><td>any</td><td>factoring</td></tr>
<tr><td>Obfuscation</td><td>poly(n)</td><td>O(1)</td><td>small</td><td>iO</td></tr>
</table>

Outline

- Local model
- Models for DP + MPC
- Lightweight architectures
  - “From HATE to LOVE MPC”
- Minimal primitives
  - “Differential Privacy via Shuffling”

This talk

Like any long, beautiful relationship,
it requires work



Your homework:

- Better protocols
- Minimal primitives
- Hybrid models
  (see A. Korolova's talk, I. Goodfellow's)
  - Nonprivate
  - Central-model DP
  - Local-model DP
- Think of other models