What Privacy Has Meant For Snips

Morten Dahl, Snips

Workshop on Privacy in Statistical Analysis, 3 May 2017

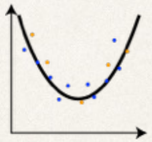
Snips

- Parisian startup
  - ~40 employees
  - majority data scientists
- On-device machine learning
  - mobile and IoT
  - strong focus on privacy
  - NLU, voice intents (B2B2C SDK)

Me

- PhD in cryptography
  - protocols and secure computation (MPC)

- at Snips
  - implementing on-device machine learning
  - experimentation and design of privacy-enhancing technologies



This Talk

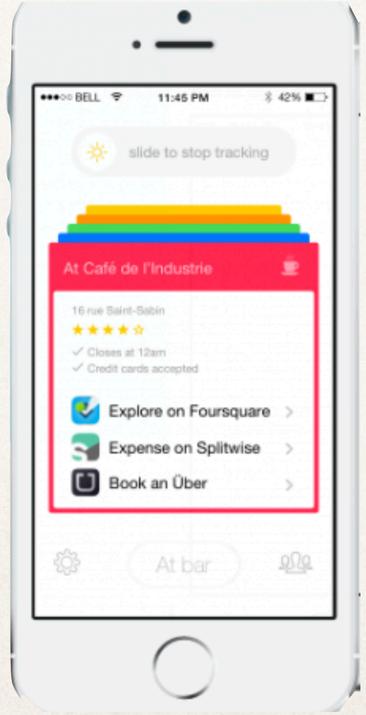
Four Encounters involving Privacy

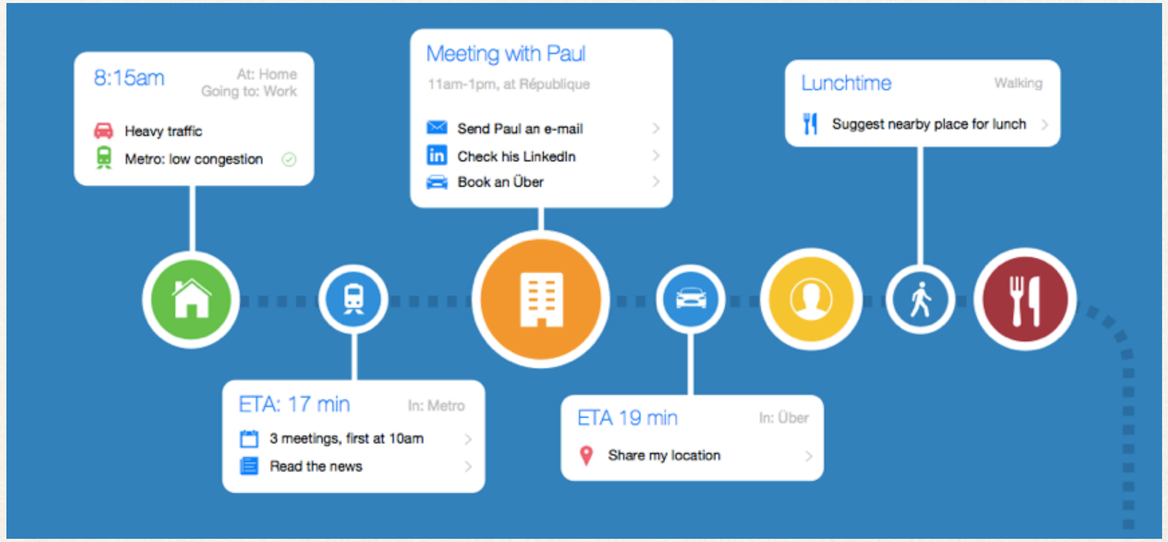
- Accessing Data
- Learning from Data
- Protecting Ourselves
- Asking for Data

Accessing Data

Context: Smart Assistant (2014)

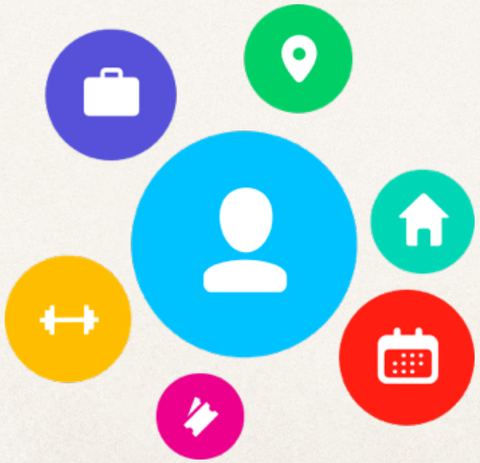
app giving personalised and context aware recommendations

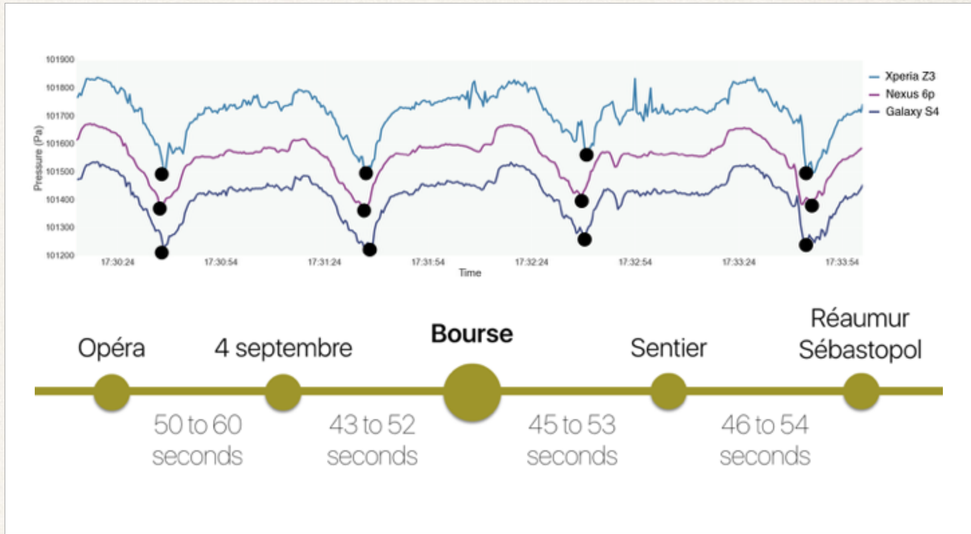




Need Data

- small company, no existing data set
  - public or generated data sets only goes so far
  - compensate by high quality sources
- data sources on mobile
  - contacts, calendar, email
  - location, activity
  - screen content*
  - barometer**





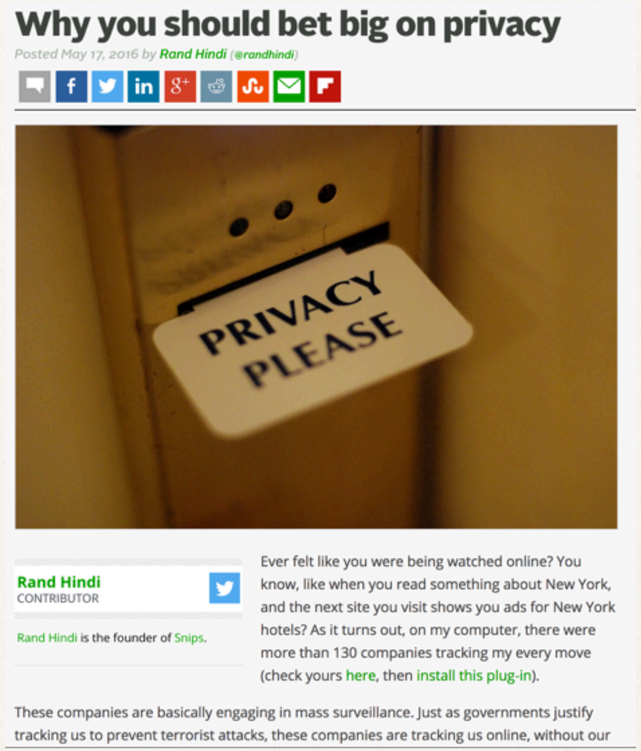
** https://medium.com/snips-ai/underground-location-tracking-3ea56803dddc

* https://medium.com/snips-ai/the-contextual-bubble-7576e4bd40c9

Getting Acceptance

- might consider this sensitive data
  - if user declines we cannot do much
  - to accept user must find the payoff worth the increased risk
  - … in this case trust that the risk is small

- how to gain trust?
  - “Privacy by Design” *
  - run everything on device
  - easy to explain to user

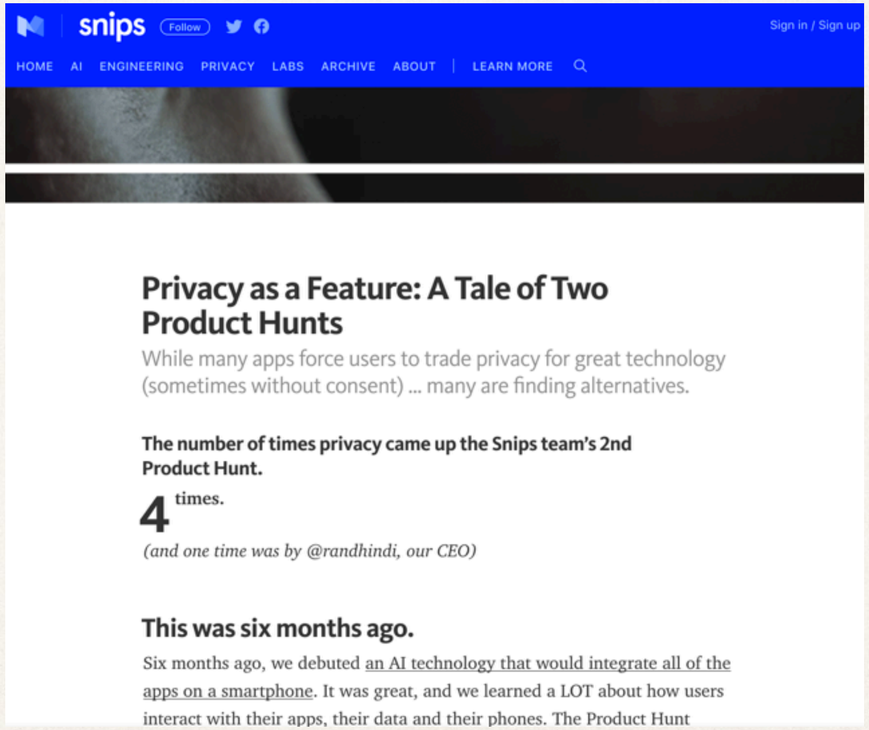


https://techcrunch.com/2016/05/17/why-you-should-bet-big-on-privacy/

* https://en.wikipedia.org/wiki/Privacy_by_design

Privacy Bonuses

- nice feeling
  - “it’s the right thing to do”
- branding differentiator
  - hiring
  - selling (funding)



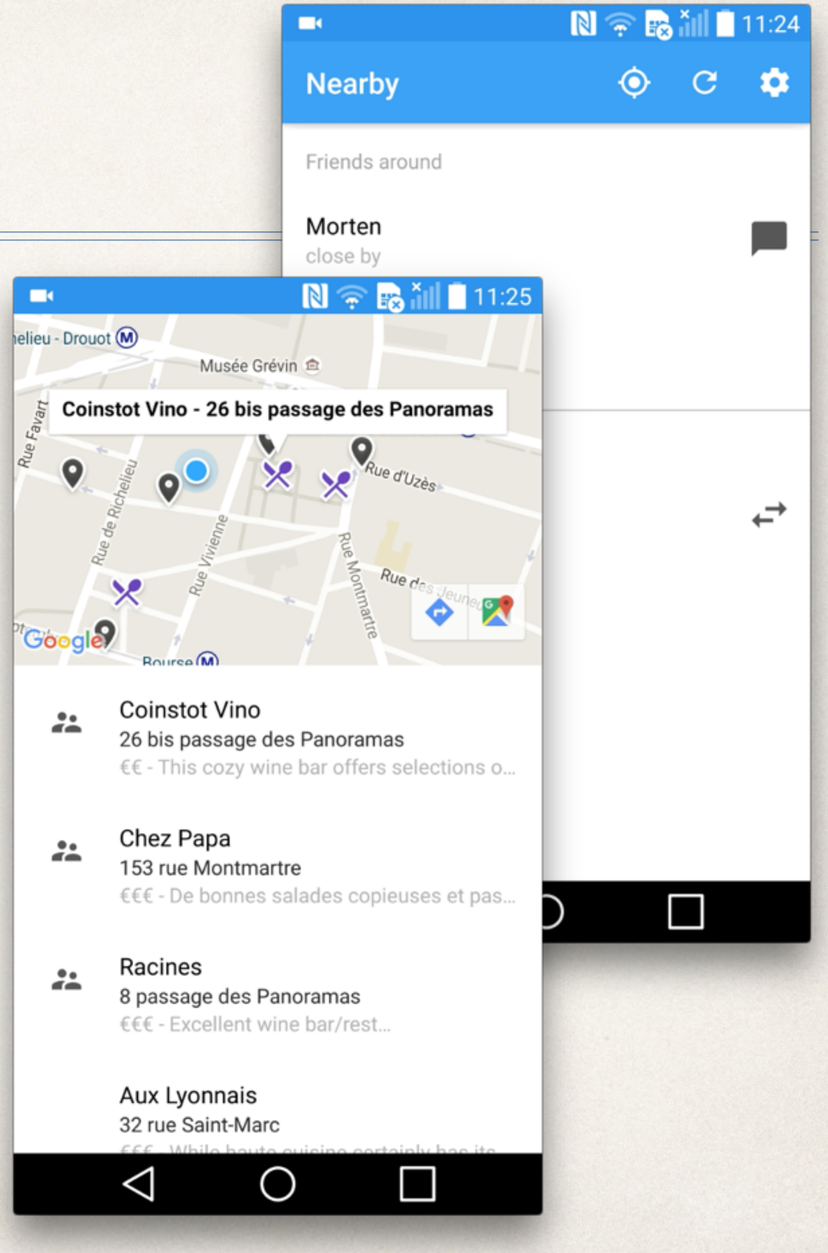
https://medium.com/snips-ai/a-tale-of-two-product-hunts-2e49c5c0f8a7

Protecting Ourselves

Context: Services (2015)

- users want to use their data
  - share: “are the people I’m about to meet close by?”
  - transform: lookup nearby places based on location

- we want users to share
  - enrich knowledge graph / context awareness
  - virality

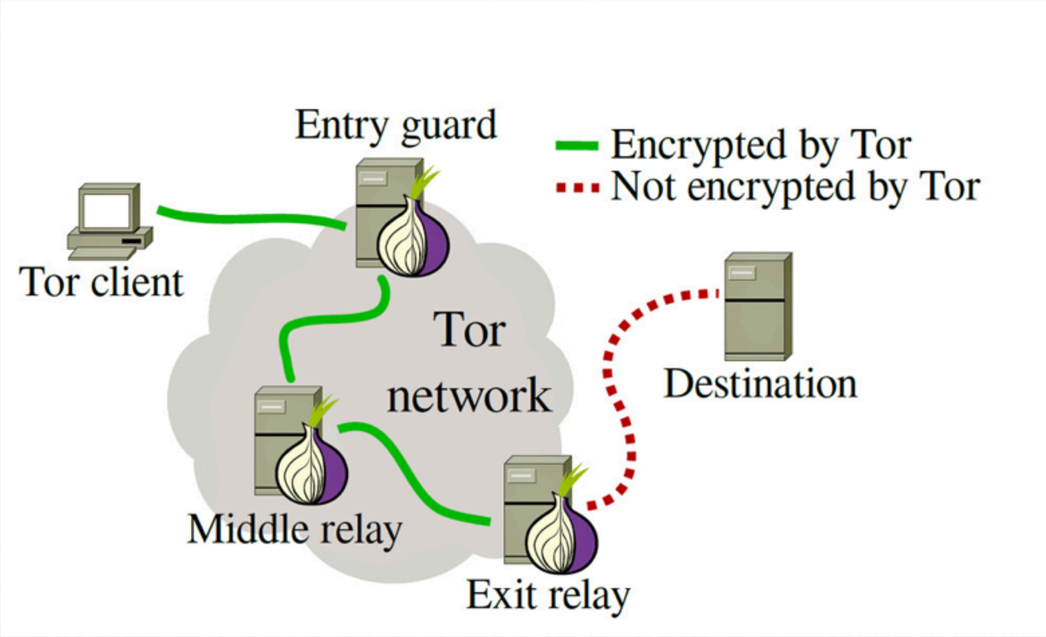


Privacy for Services

- surprises would hurt trust
  - explicit and opt-out
  - bonus: continue to work offline
  - high acceptance rate
- technical solutions?
  - first look at cryptography

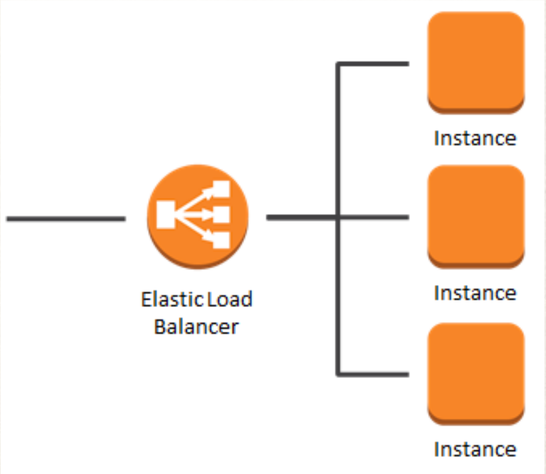
Hide user IP address

- remove obvious identifiers
  - IP address links all requests together
- Tor network as proxy
  - user cost high (mobile setting)
  - active attacker
- statistical attacks to re-identify users



What are we protecting against?

- we're generally good guys
  - (still need to convince user of course)
- we cannot overspend on security
  - likely break-in one day
  - will hurt trust
  - "Data is a Toxic Asset" *
  - protect against accidental data storage
- load balancer on Amazon AWS as proxy
  - user cost low
  - administrator could bypass
  - reduce places we can mess up



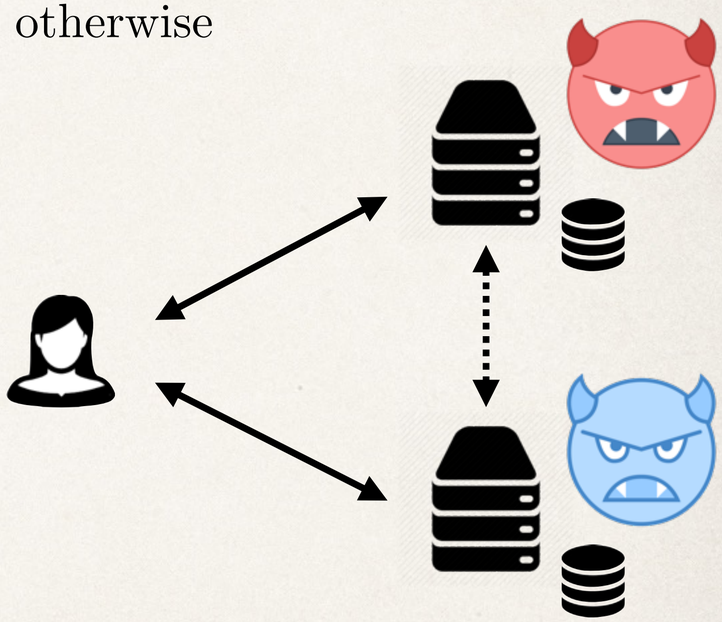
* https://www.schneier.com/blog/archives/2016/03/data_is_a_toxic.html

Hide user content

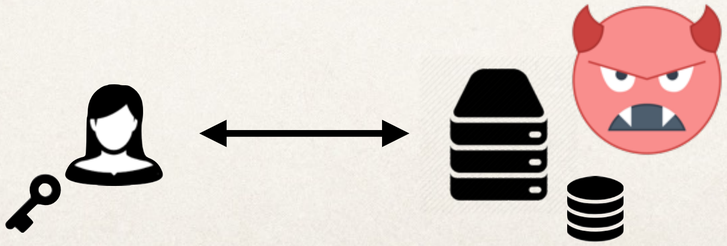
- Private Information Retrieval
  - privately evaluate point functions without revealing point

$$f_p(x) = \begin{cases} 1 & \text{if } x = p \\ 0 & \text{otherwise} \end{cases}$$

- two flavours
  - using secret sharing (several servers)
  - using homomorphic encryption (single server)



- experimental implementations
  - XPIR [ABFK'14]



Homomorphic Encryption

Partially HE	Somewhat HE	Fully HE
Enc(x) + Enc(y) = Enc(x + y)	…	…
Enc(x) * y = Enc(x * y)	Enc(x) * Enc(y) = Enc(x * y)	Enc(x) * Enc(y) = Enc(x * y)
	(bounded number of times)	*(unbounded)*

can be one or more ciphertexts

size and efficiency depends heavily on encoding

Enc(index) →

(index1 == Enc(index)) * area1
+ (index2 == Enc(index)) * area2
+ (index3 == Enc(index)) * area3
= Enc(area)

← Enc(area)

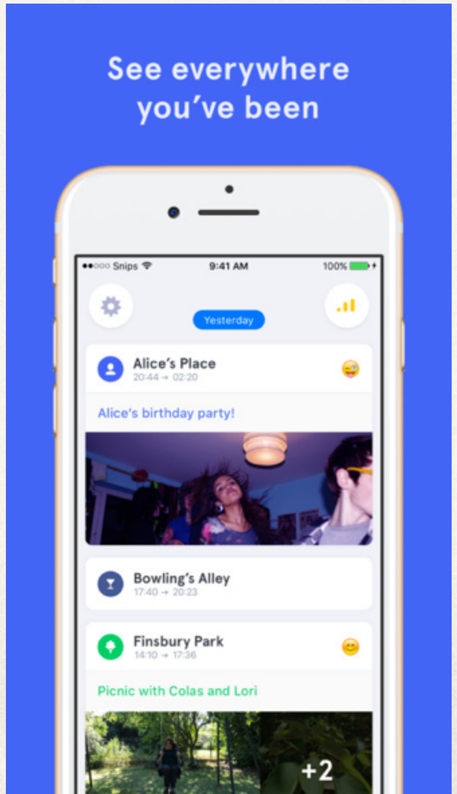
Experiments for Paris: ~3000 buckets

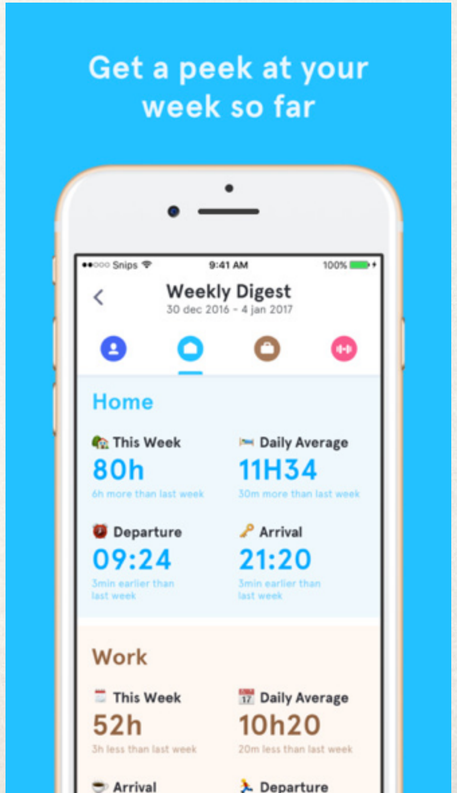
- relatively computationally efficient
  - processing < 1sec
- large query size
  - upload + download < 300KB
  - better to download entire database for repeating queries
- best for few queries and large responses

Learning from Data

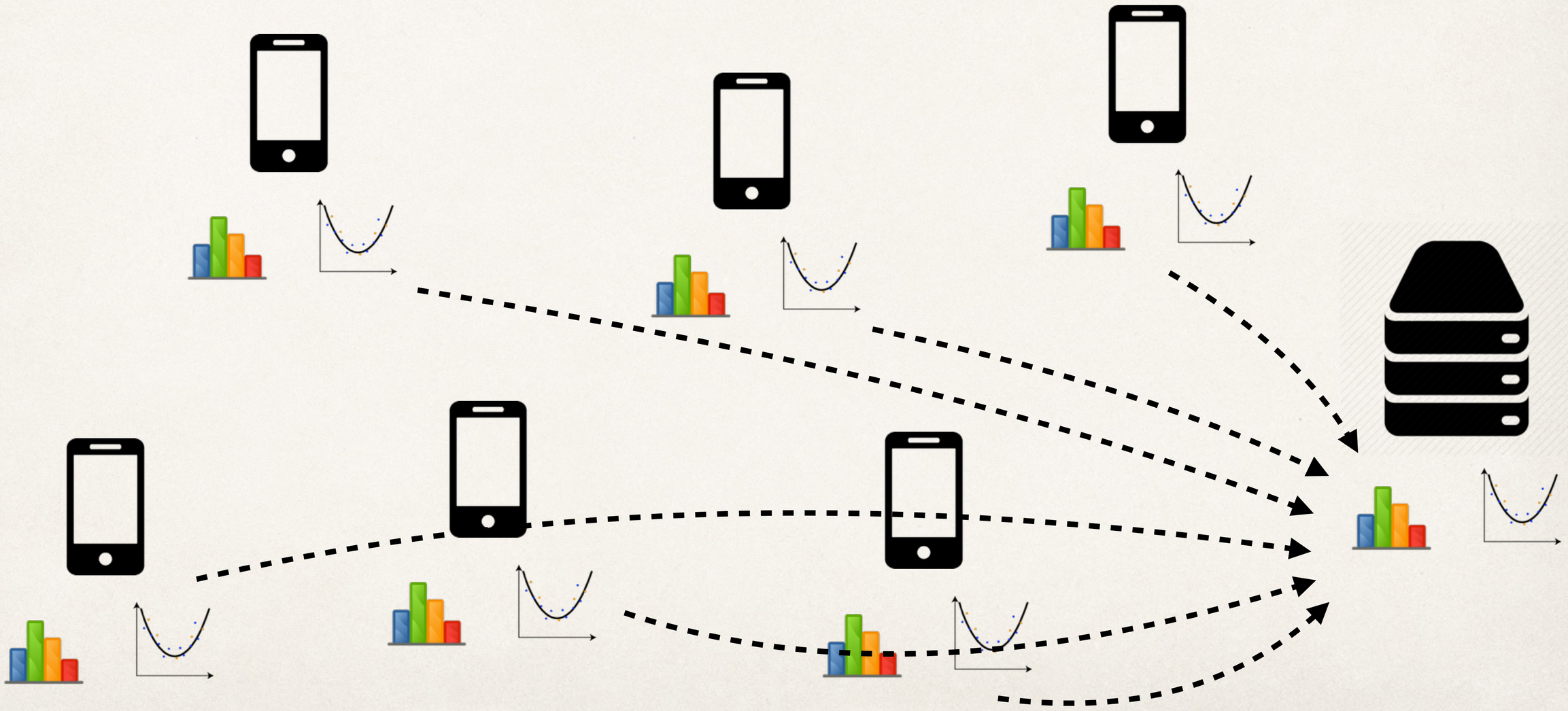
Context: SDK and Capsule (2016)

- recommendations
  - local data not always interesting
    - first-launch, travelling, …
  - manual generation of prior doesn't scale
  - use similar users to give better experience
- experiments
  - opt-in
  - harder to explain: blog posts, academia, open source



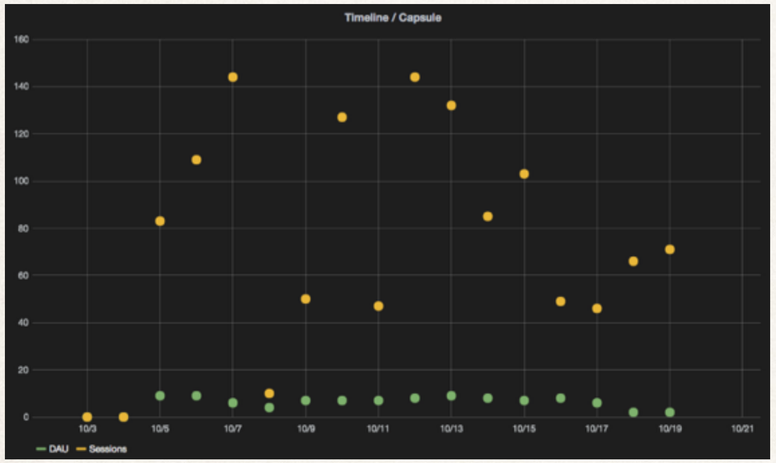


Learning from Distributed Data Sets

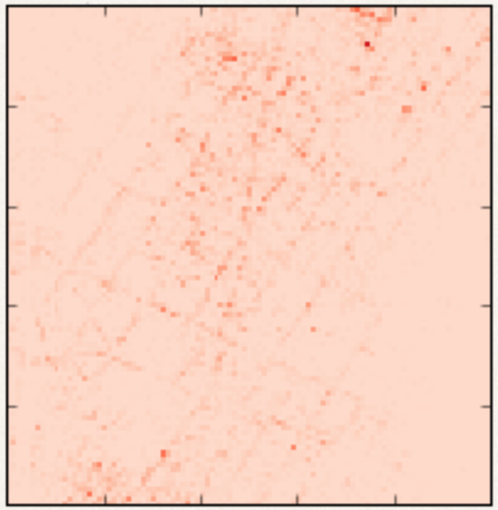


Aggregation of Distributed Data Sets

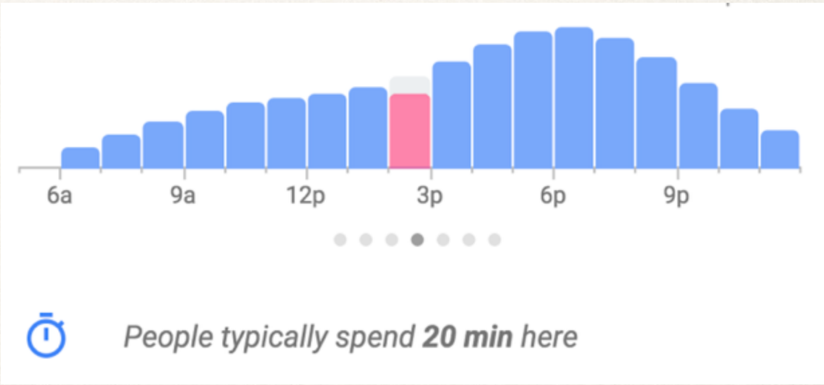
analytics / surveys



discovery



recommendations



$$x = \sum x_i$$

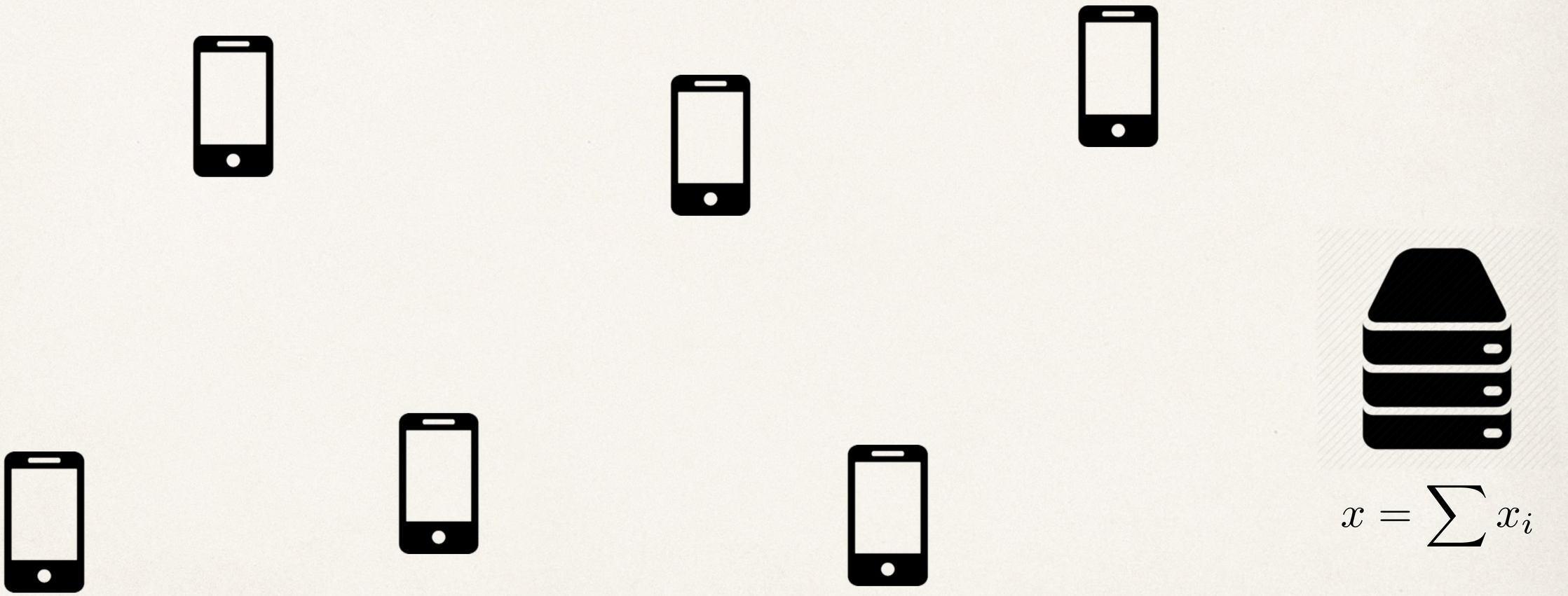
$$dim(x_i) \geq 10k$$

Constraints

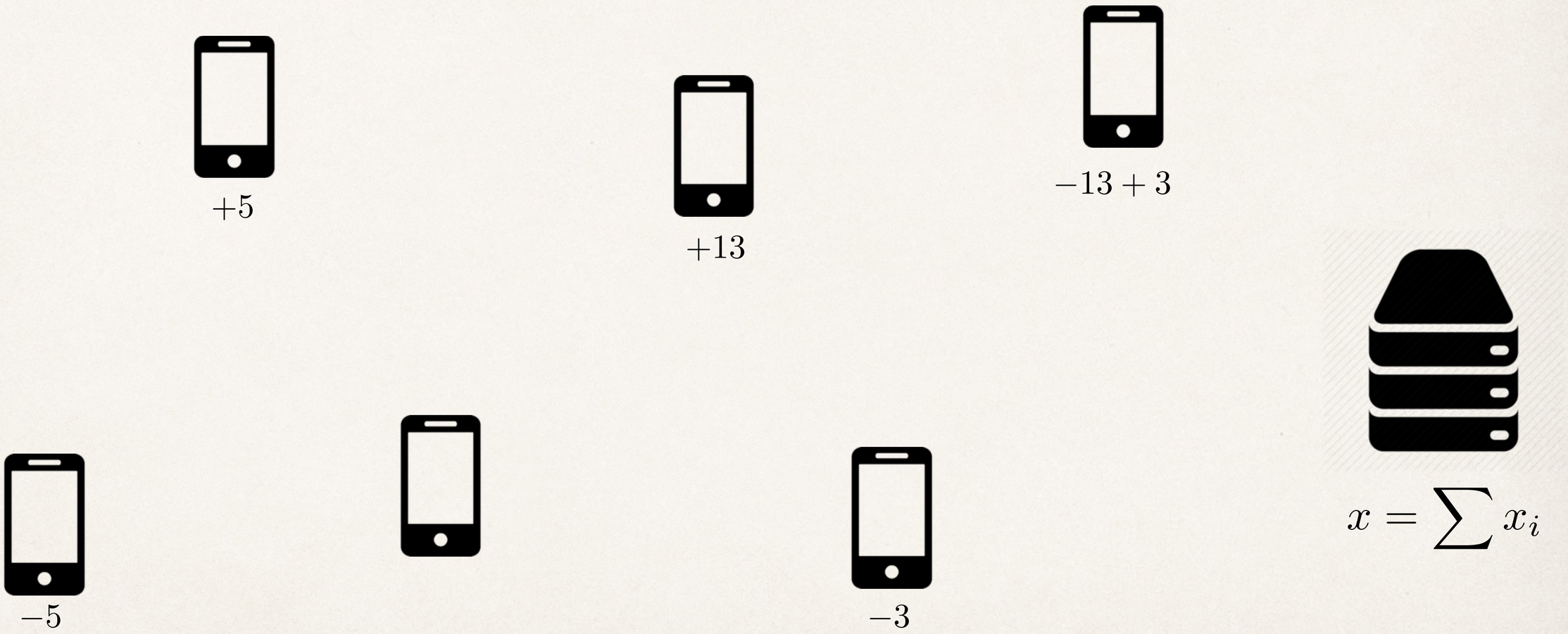
no individual user data

Mobile and Web

limited computation	limited connectivity	sporadic behaviour
minimise device processing	minimise session count + length	minimise coordination


$x = \sum x_i$

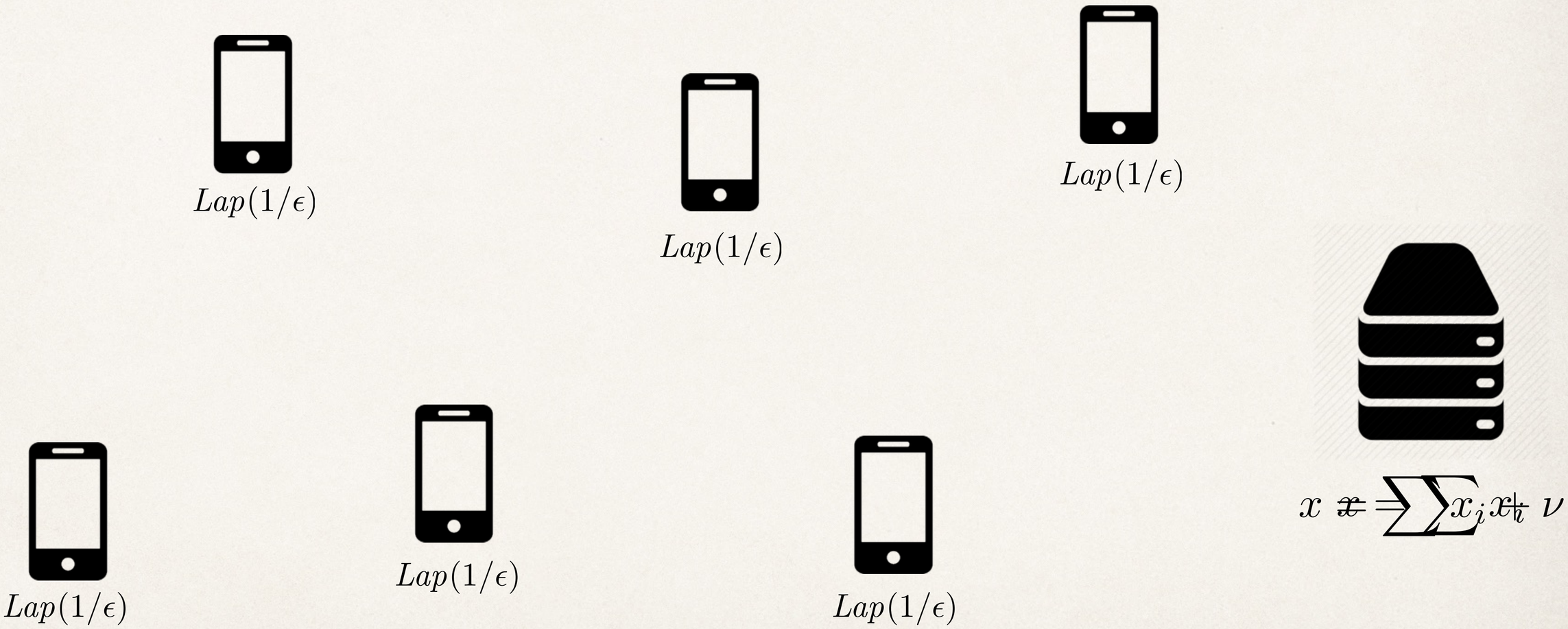
Sensor Networks



high performance

coordination

(Local) Differential Privacy

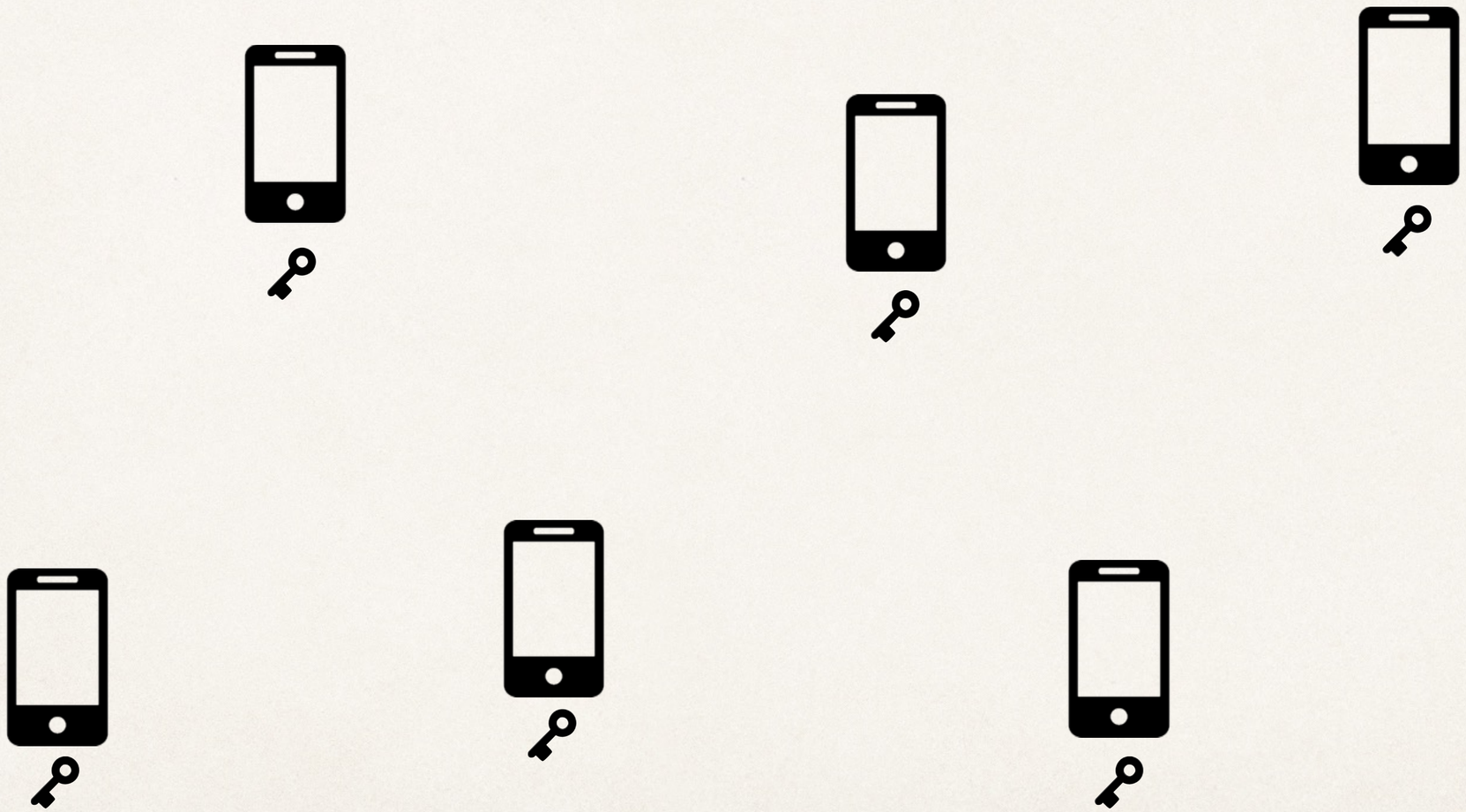


high performance

formal privacy

signal-to-noise

Homomorphic Encryption



$$x = \sum x_i$$

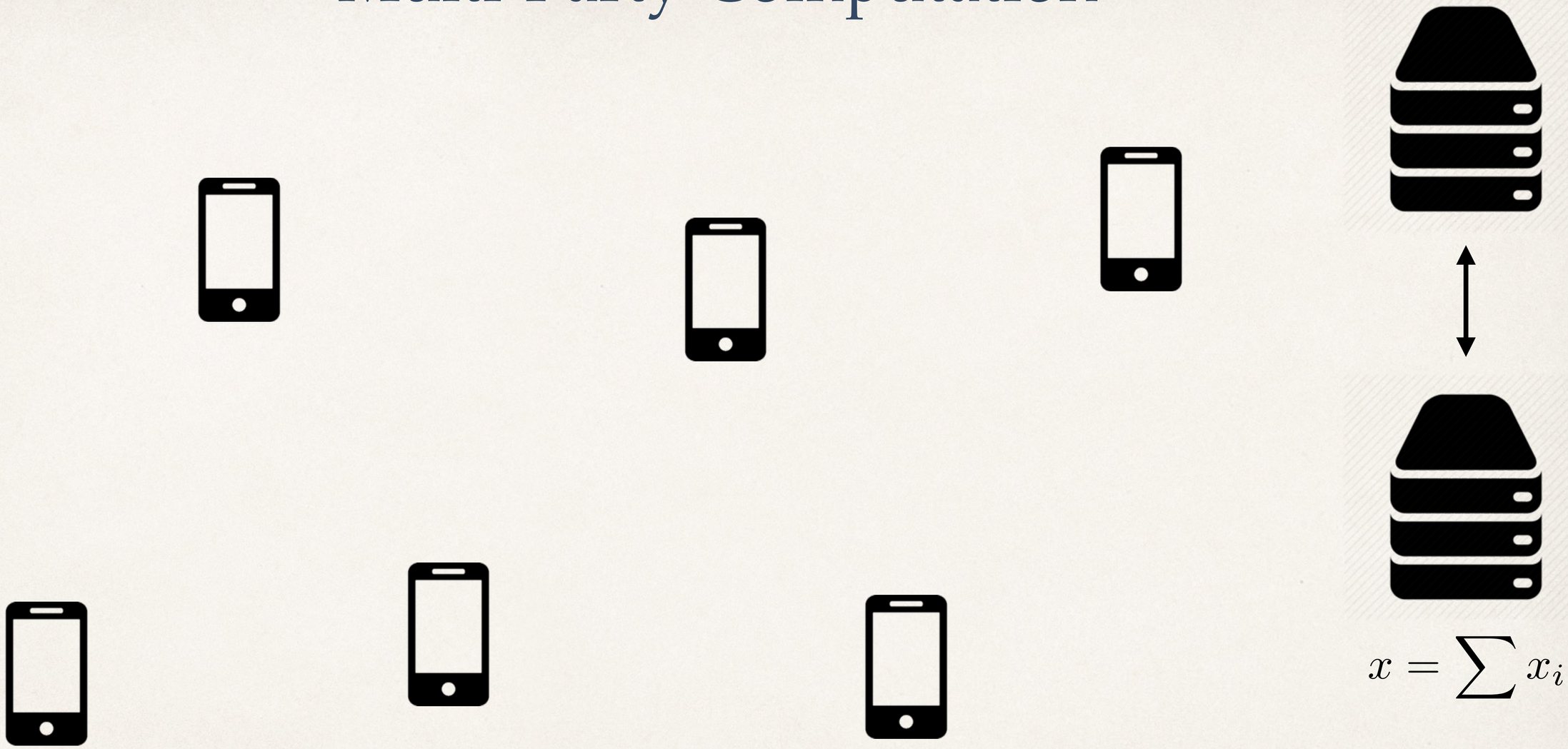
flexible

explainable

efficiency

keys

Multi-Party Computation



flexible

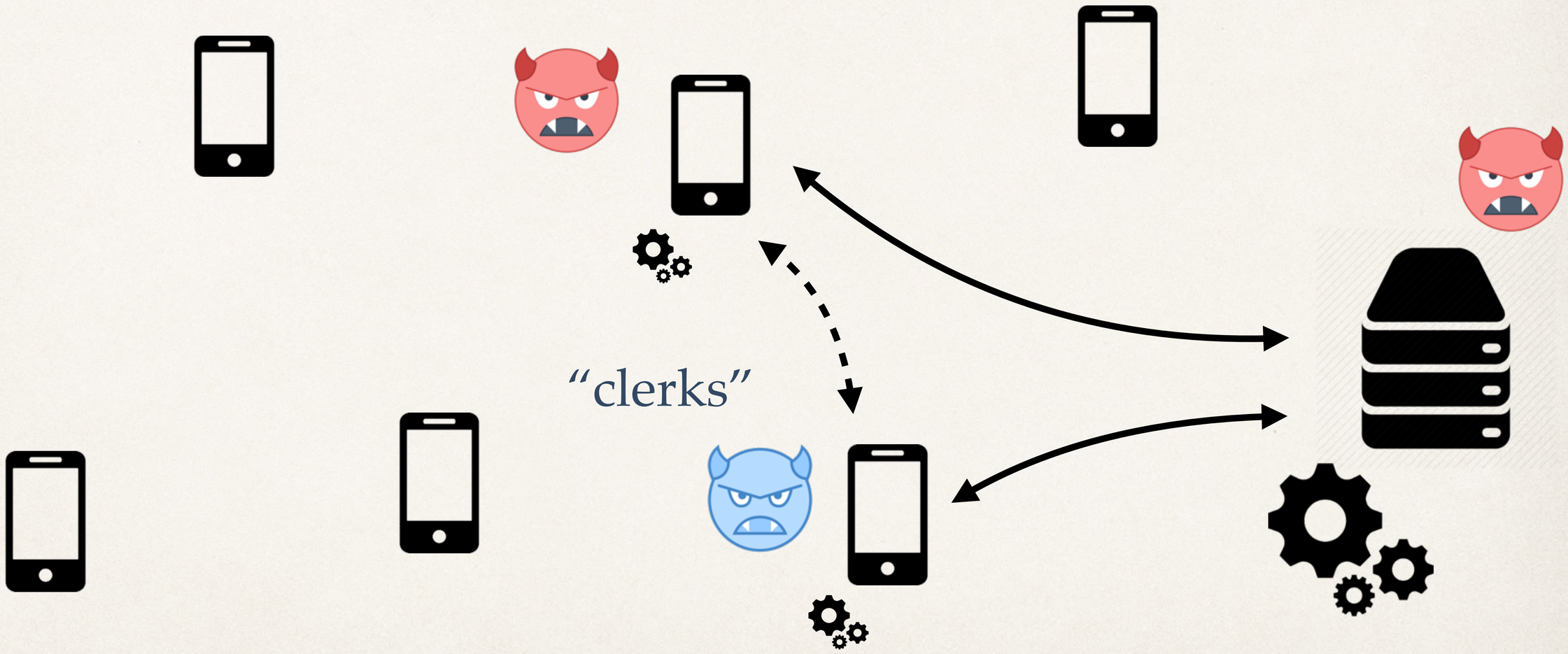
decent performance

interaction

no one to play with

Community Trust

addressing the problem of only one powerful server



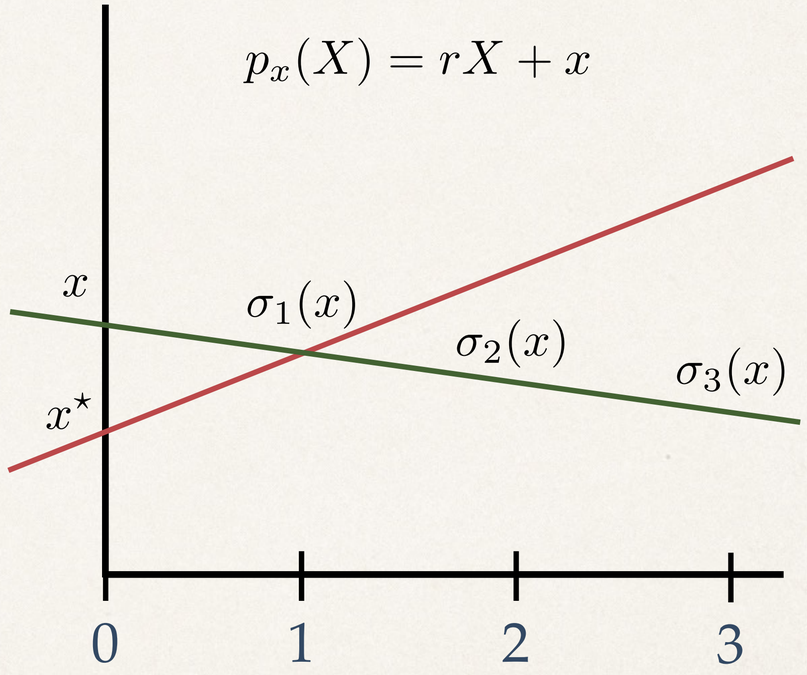
Homomorphic Secret Sharing

$$\text{Share}(x) = \Big(\sigma_1(x), \sigma_2(x), \sigma_3(x)\Big)$$

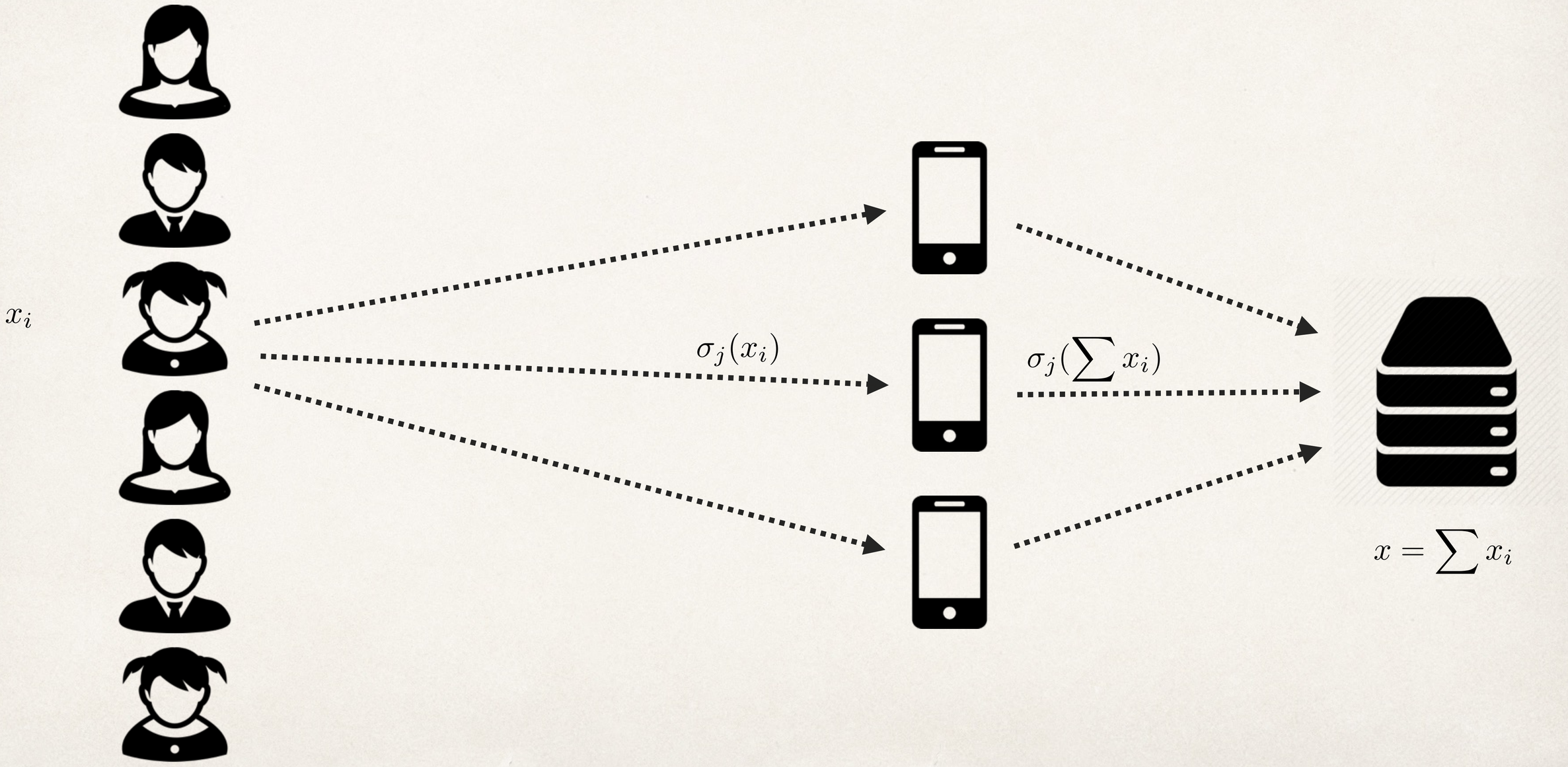
$$\text{Reconstruct}\Big(\sigma_1(x), \sigma_2(x), \sigma_3(x)\Big) = x$$

$$deg(p_x) = t \qquad p_x(0) = x$$

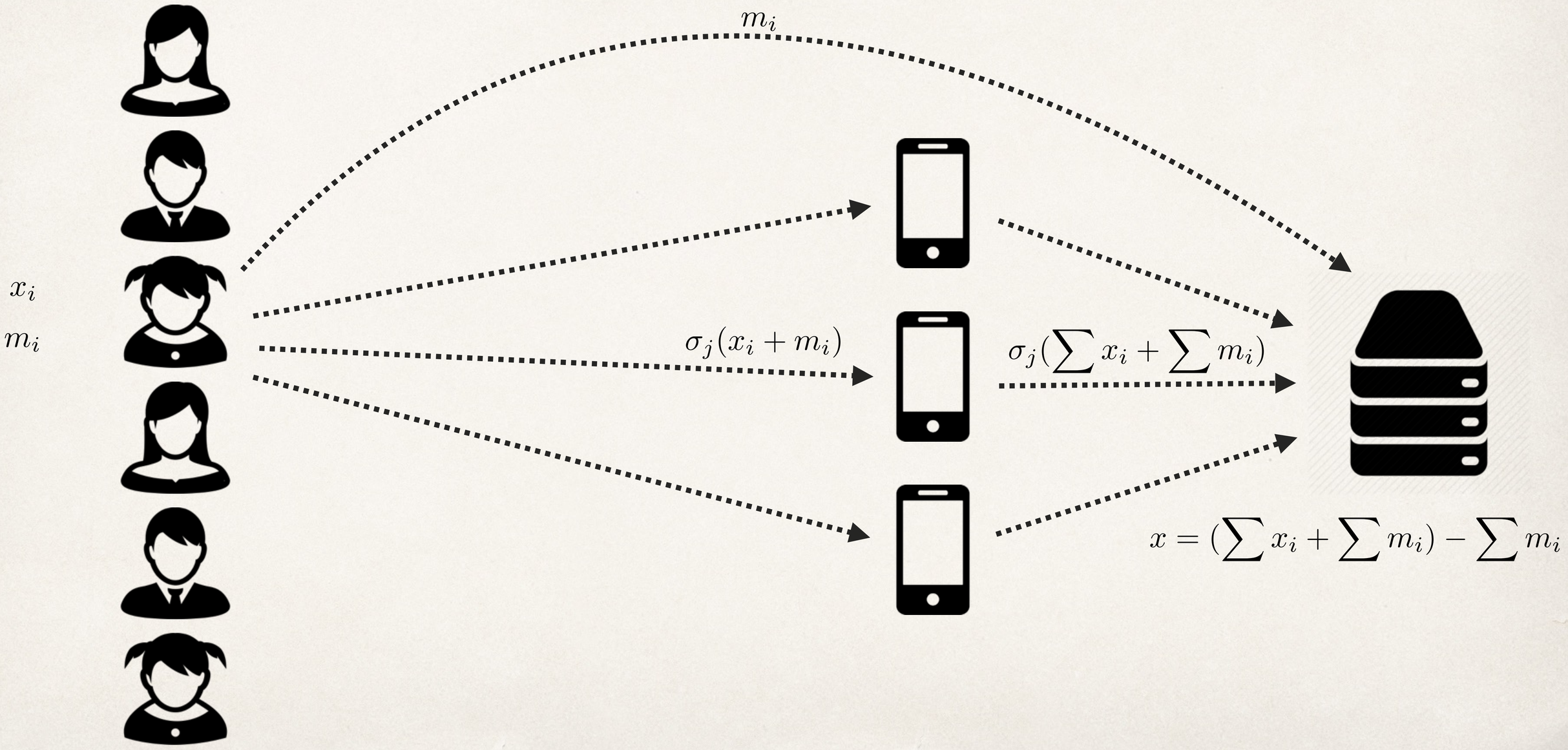
$$\sigma_j(x) = p_x(j)$$



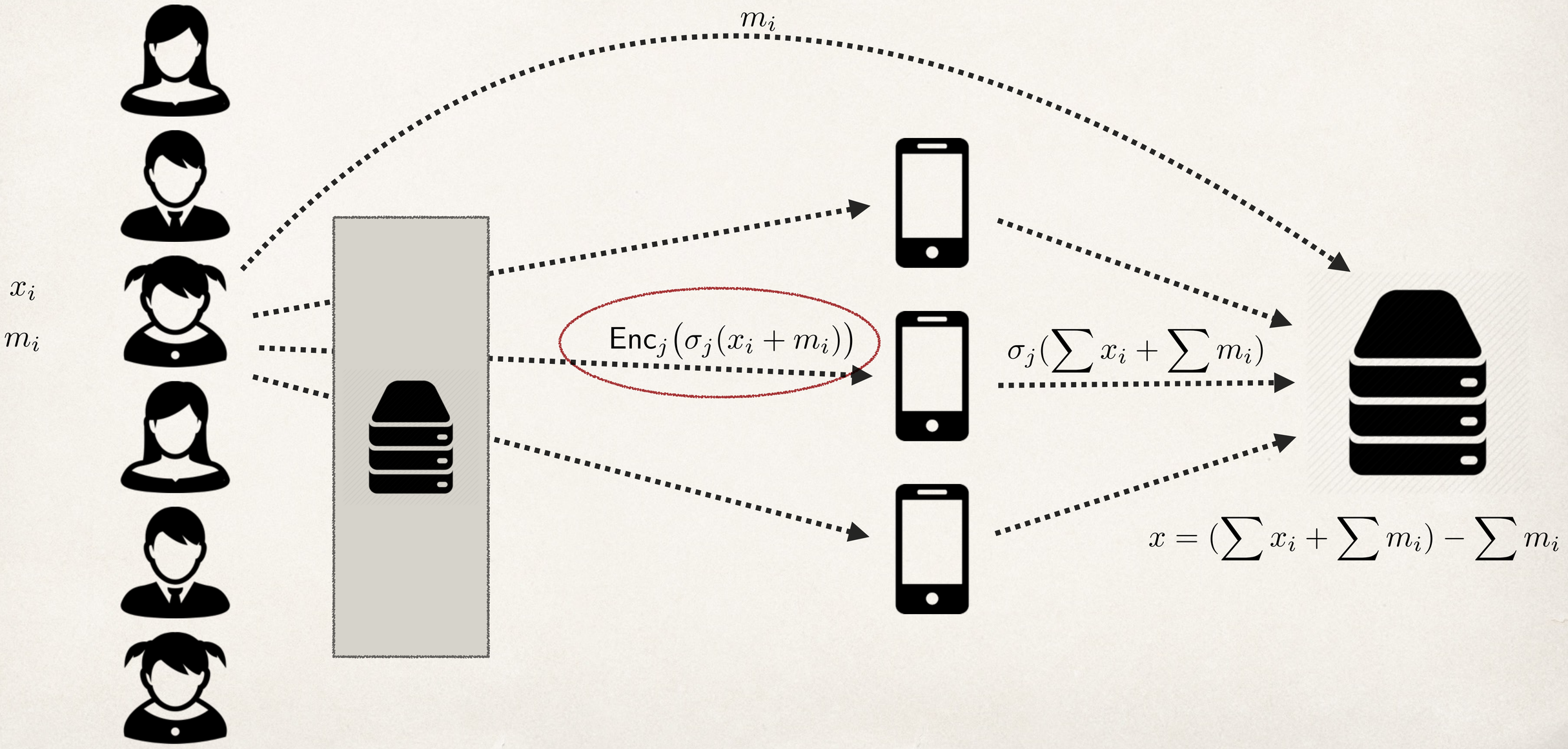
Protocol



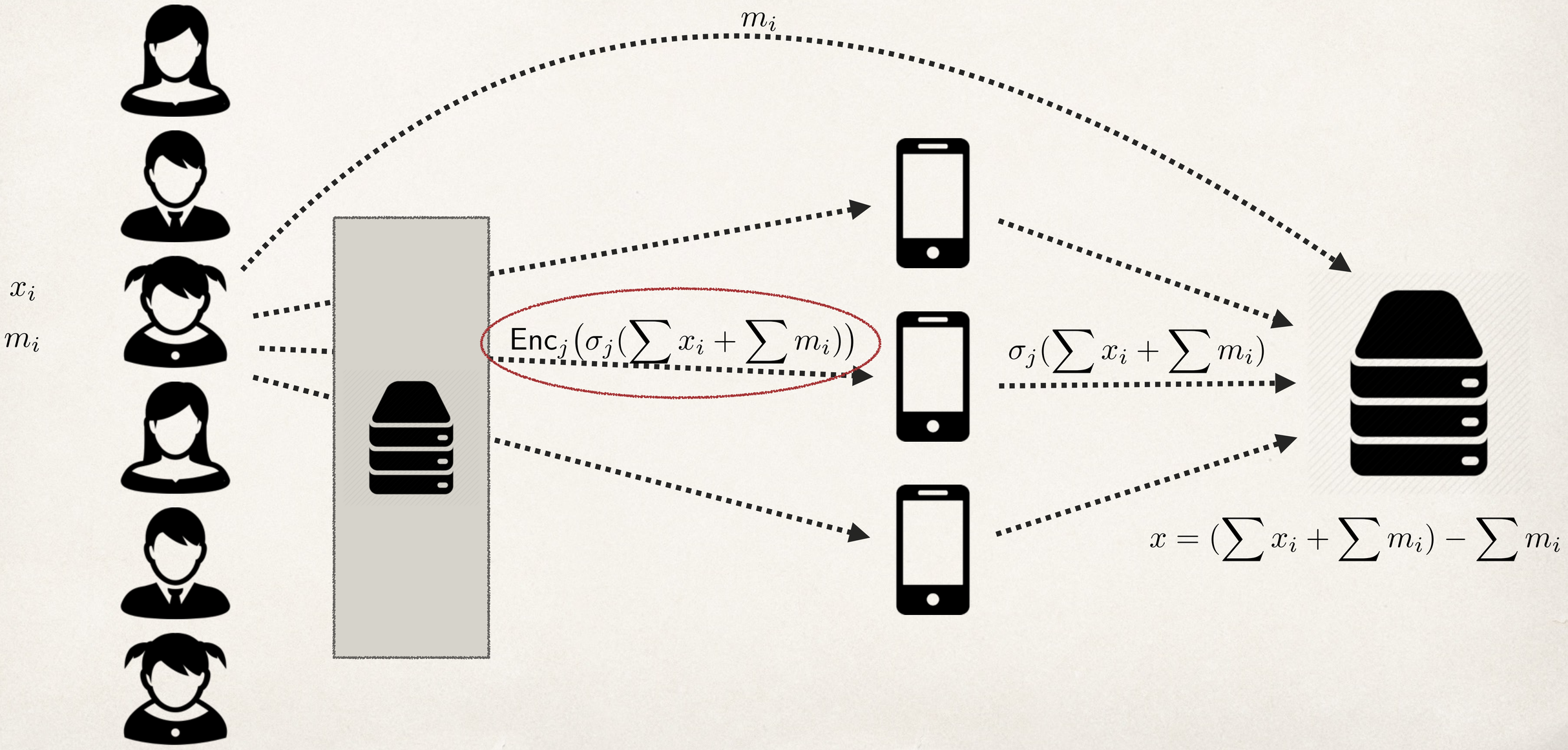
Protocol



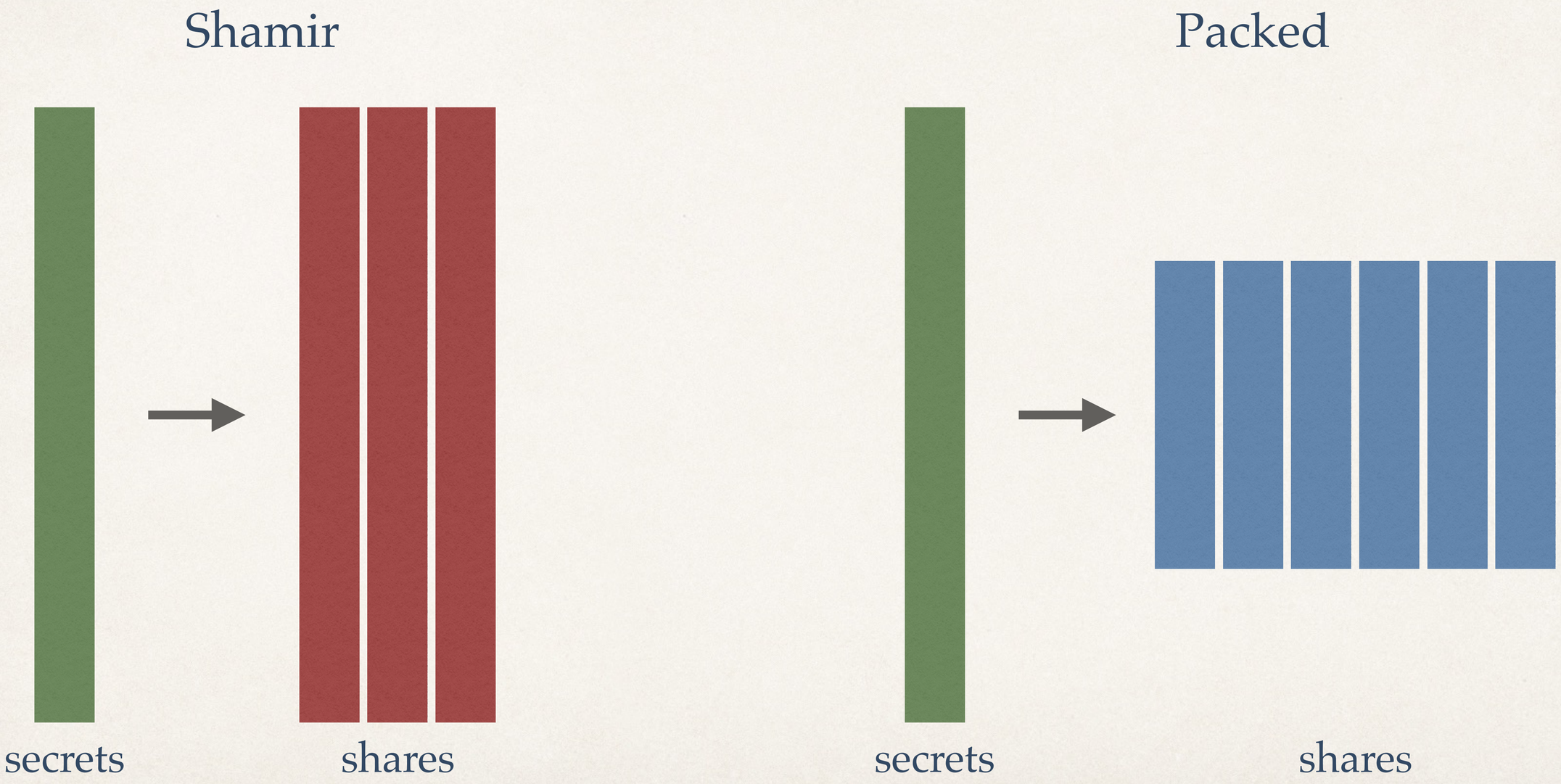
Protocol


m_i
x_i
m_i
$\mathsf{Enc}_j\big(\sigma_j(x_i + m_i)\big)$
$\sigma_j(\sum x_i + \sum m_i)$
$x = (\sum x_i + \sum m_i) - \sum m_i$

Protocol


m_i
x_i
m_i
$\mathsf{Enc}_j(\sigma_j(\sum x_i + \sum m_i))$
$\sigma_j(\sum x_i + \sum m_i)$
$x = (\sum x_i + \sum m_i) - \sum m_i$

Distributing the Workload

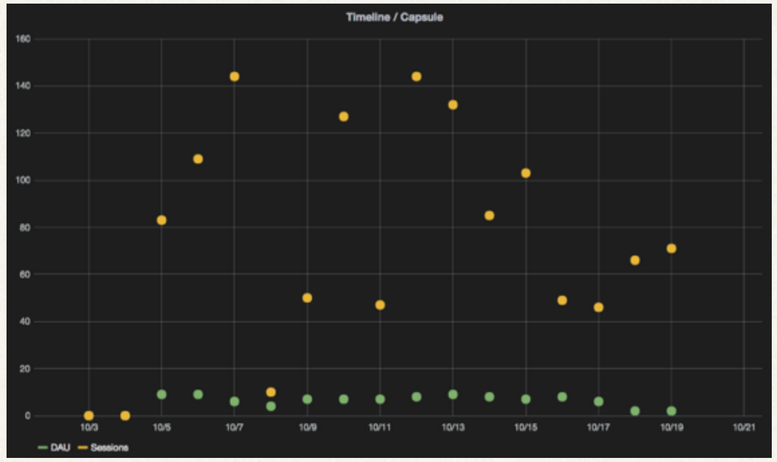


Result

lightweight MPC protocol for linear functions,
tailored for large-scale high-dimension aggregation

Users	Clerks	Server	DP
single message	easy setup	most of the work	one extra round
	work distribution		
	level of resilience		
	some active security		

Analytics



Communication

dimension 100

non-homomorphic

download	small	medium	large
25 000	977KB		
80 000		938KB	
250 000			977KB

Averages

Computation

dimension 35k

homomorphic

unlimited users

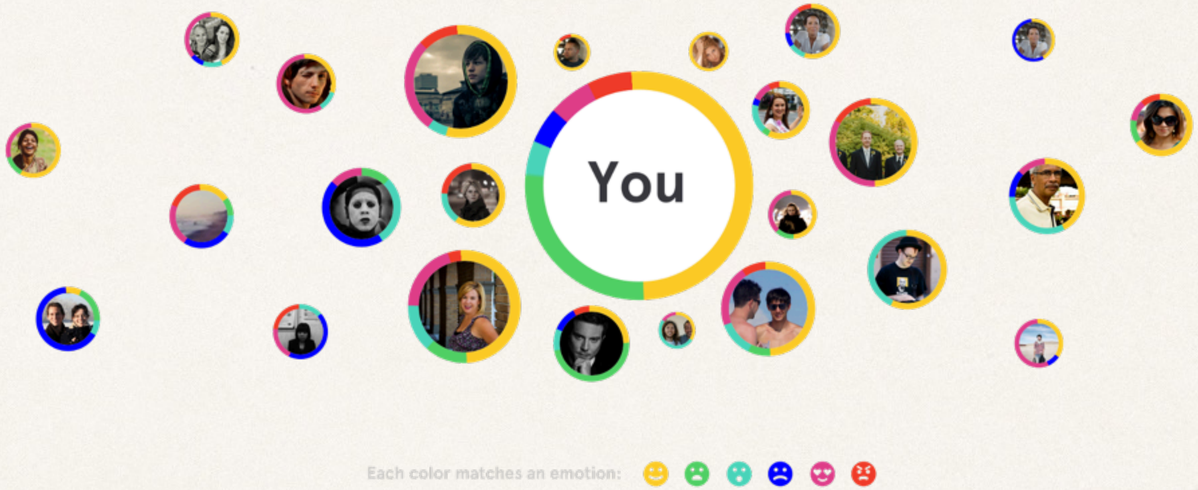
decrypt	small	medium	large
RP	21.6s	4.6s	0.6s
iPhone	6.2s	1.3s	0.2s
MBP	0.9s	0.2s	0.03s

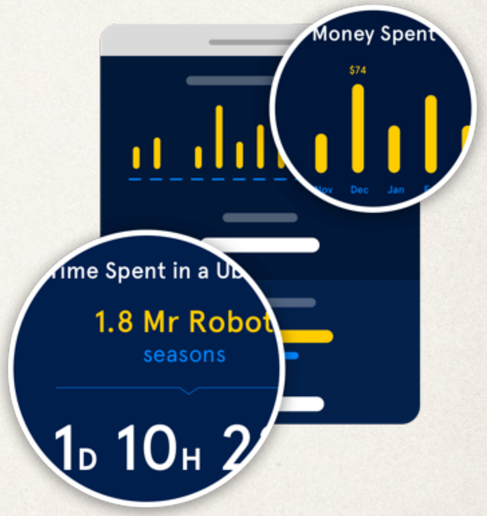
share	small	medium	large
RP	0.3s	0.2s	0.2s
iPhone	0.3s	0.3s	0.3s
MBP	0.06s	0.03s	0.03s

Asking for Data

Context: NLU for Chat

- corpus for chat messages
  - entity recognition
- high dimension
  - how to formalise privacy?
  - cryptographic techniques unfit
- resort to anonymization
  - opt-in
  - remove (identified) sensitive entities
  - transparency: user review and filter before sending
  - give something in return

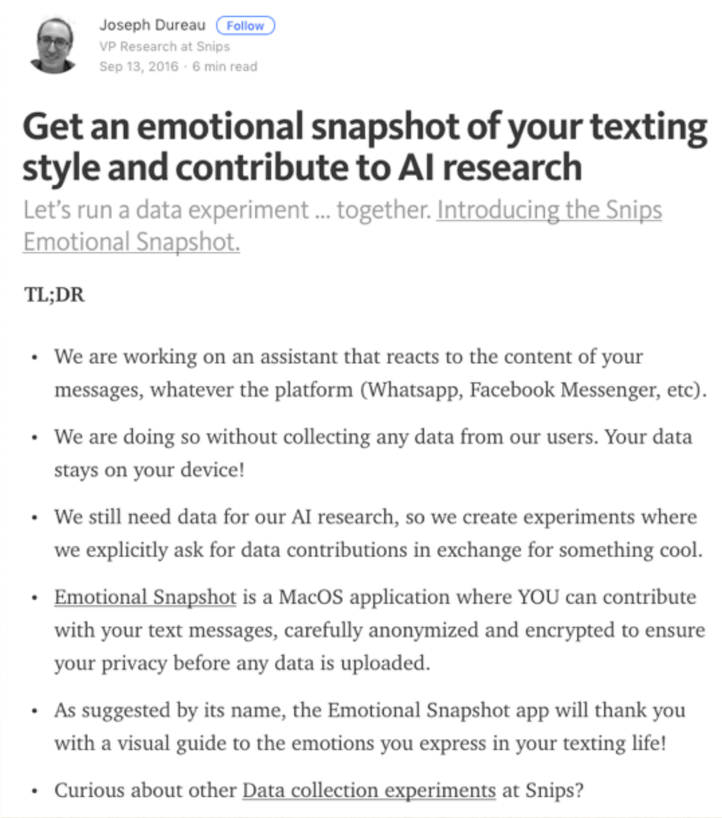




https://snips.ai/seeds/

Heuristics

- anonymization process
  - remove phone numbers, credit cards, social security numbers
  - remove received chat messages
  - round timestamps
- limit exposure
  - protected server
  - mix messages
  - filter what's not related to current study
  - each data scientist (mechanical turk) only sees a few



Joseph Dureau Follow
VP Research at Snips
Sep 13, 2016 · 6 min read

Get an emotional snapshot of your texting style and contribute to AI research

Let's run a data experiment ... together. Introducing the Snips Emotional Snapshot.

TL;DR

- We are working on an assistant that reacts to the content of your messages, whatever the platform (Whatsapp, Facebook Messenger, etc).
- We are doing so without collecting any data from our users. Your data stays on your device!
- We still need data for our AI research, so we create experiments where we explicitly ask for data contributions in exchange for something cool.
- Emotional Snapshot is a MacOS application where YOU can contribute with your text messages, carefully anonymized and encrypted to ensure your privacy before any data is uploaded.
- As suggested by its name, the Emotional Snapshot app will thank you with a visual guide to the emotions you express in your texting life!
- Curious about other Data collection experiments at Snips?

https://medium.com/snips-ai/step-into-the-lab-with-snips-lets-run-a-data-experiment-together-b12a902137ea

Summary

This Talk

- Accessing Data
  - gain trust
  - run locally
  - easy to explain
- Protecting Ourselves
  - risk management
  - Data as a Toxic
  - prevent accidental storage
- Learning from Data
  - mixing cryptographic tools
  - hard to explain
  - trust from community
- Asking for Data
  - hard to formalise
  - give something in return
  - users willing to share

Tools

DP

- hard to approach
- parameter choices
- implementations

MPC

- maturing
- other servers?

1.3M AES
(7B gates)
per second

https://eprint.iacr.org/2016/768

HE

- behind MPC
- several users?

9000 adds
1500 muls
per second

Louis/Pedro

On-device

- easy to explain
- low latency
- no servers to pay for

Additional Notes

Data Breach Insurance *

- lower cost
- limited coverage

NDAs **

- established ways of dealing with breach
- shifts explanation burden

Regulation ***

- 20M euro liability

* https://www.business.com/articles/do-you-need-data-breach-insurance/

** www.cs.bris.ac.uk/Research/CryptographySecurity/TPMPC/Slides2017/MayankVaria.pptx

*** https://blog.lukaszolejnik.com/interesting-points-in-leak-of-new-eprivacy-directive/

Thank you!

@mortendahlcs

mortendahlcs@gmail.com

@snips

medium.com/snips-ai