Differential Privacy Techniques Beyond Differential Privacy

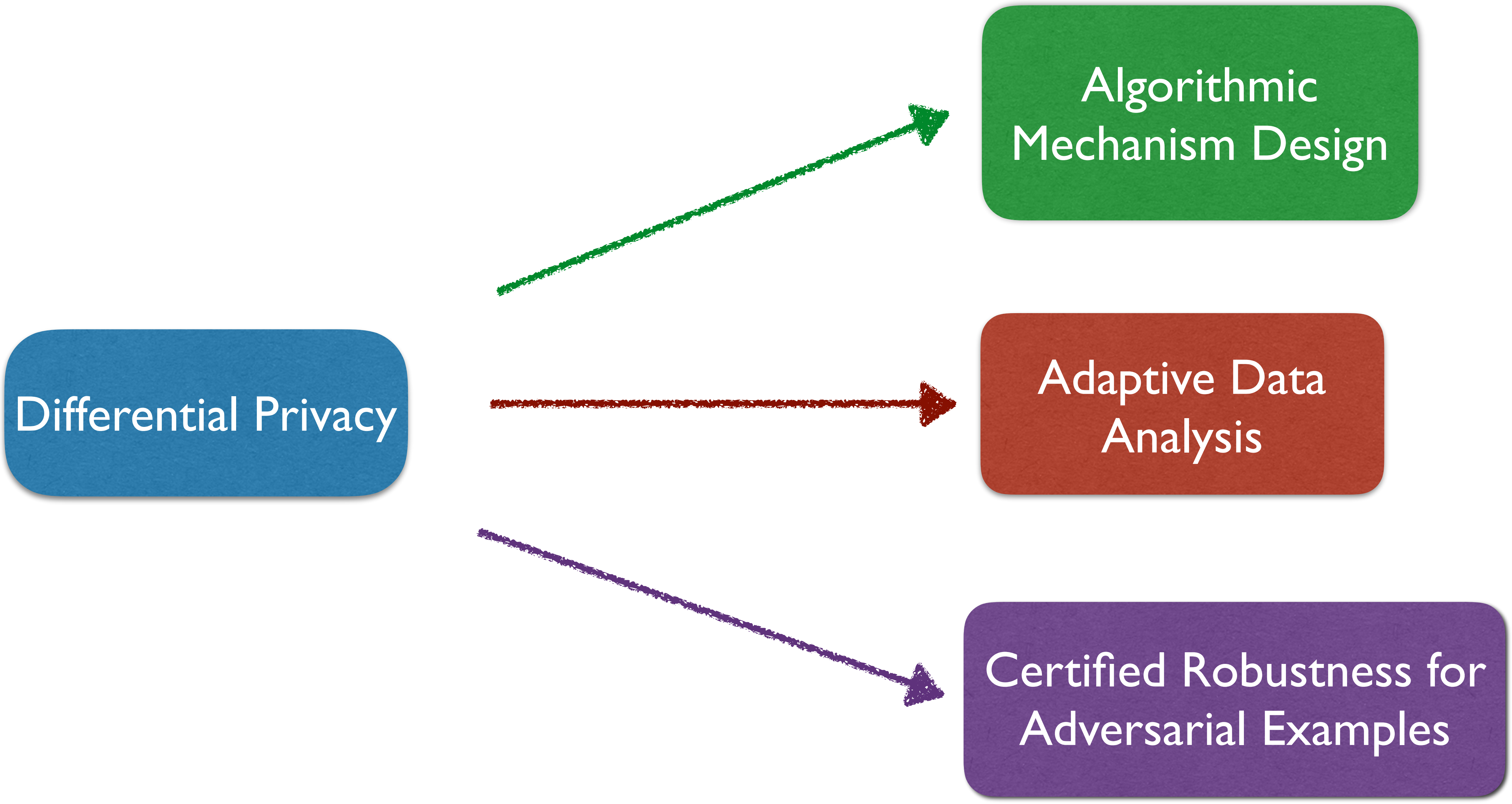
Steven Wu
Assistant Professor
University of Minnesota

"Differential privacy? Isn't it just adding noise?"

How to add smart noise to guarantee privacy without sacrificing utility in private data analysis?

How to add smart noise to achieve stability and gain more utility in data analysis?!

Technical Connections



Outline

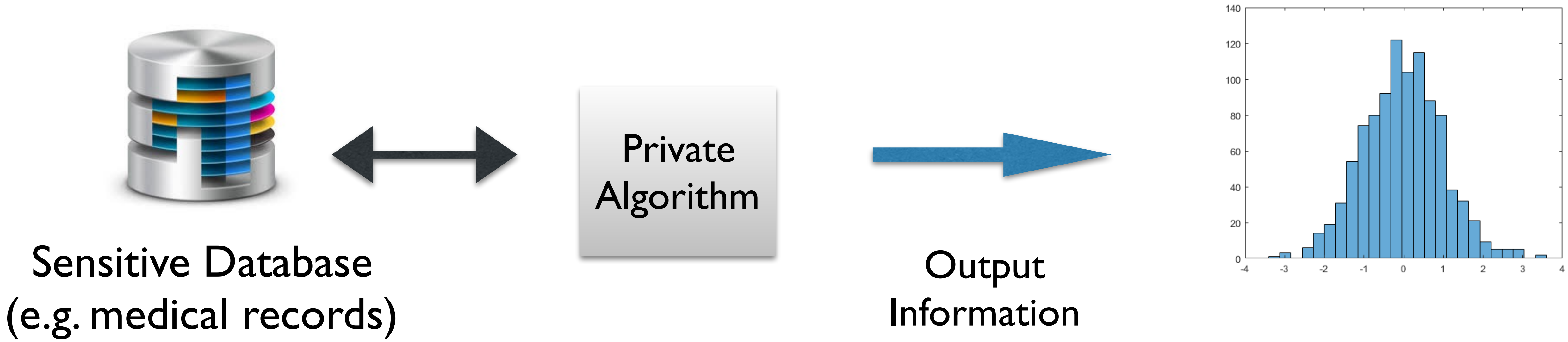
- Simple Introduction to Differential Privacy
- Mechanism Design
- Adaptive Data Analysis
- Certified Robustness

Outline

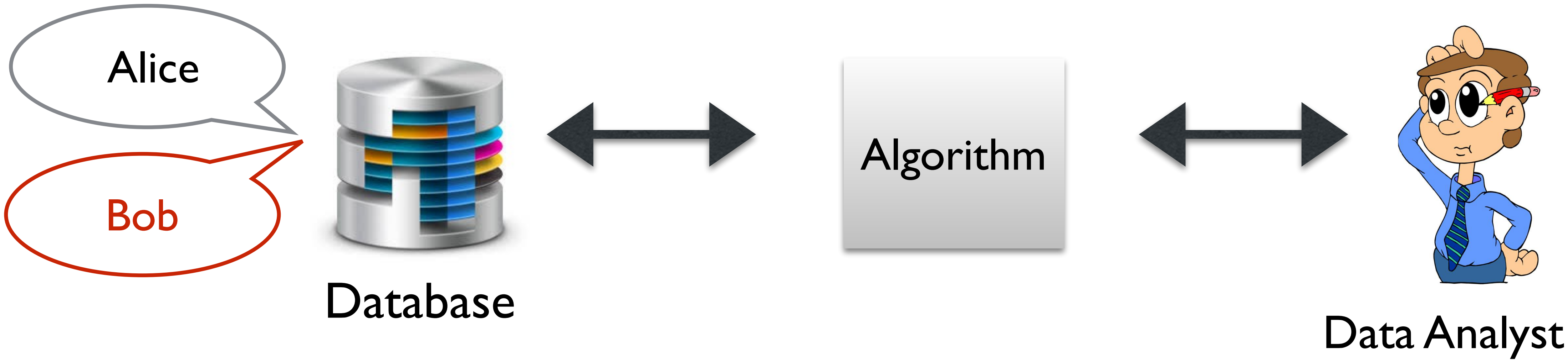
- Simple Introduction to Differential Privacy
- Mechanism Design
- Adaptive Data Analysis
- Certified Robustness

Statistical Database

- X: the set of all possible records (e.g. $\{0, 1\}^d$)
- $D \in X^n$: a collection of n rows ("one row per person")



Privacy as a Stability Notion



Stability: the data analyst learns (approximately) same information if any row is replaced by another person of the population

Differential Privacy

[DN03, DMNS06]

D =

D_1
D_2
D_3
…
D_n

D' =

D_1
D_2
D'_3
…
D_n

D and *D'* are *neighbors* if they differ by at most one row

A private algorithm needs to have close output distributions on any pair of neighbors

Definition: A (randomized) algorithm *A* is ε-differentially private if for all neighbors *D*, *D'* and every S ⊆ Range(A)

$$\Pr[A(D) \in S] \leq e^{\varepsilon} \Pr[A(D') \in S]$$

Differential Privacy

[DN03, DMNS06]

Definition: A (randomized) algorithm *A* is (ε, δ)-differentially private if for all neighbors *D*, *D'* and every S ⊆ Range(A)

$$\Pr[A(D) \in S] \leq e^{\varepsilon} \Pr[A(D') \in S] + \delta$$

One Interpretation of the Definition:

If a bad event is very unlikely when I'm not in the database (*D*), then it is still very unlikely when I am in the database (*D'*).

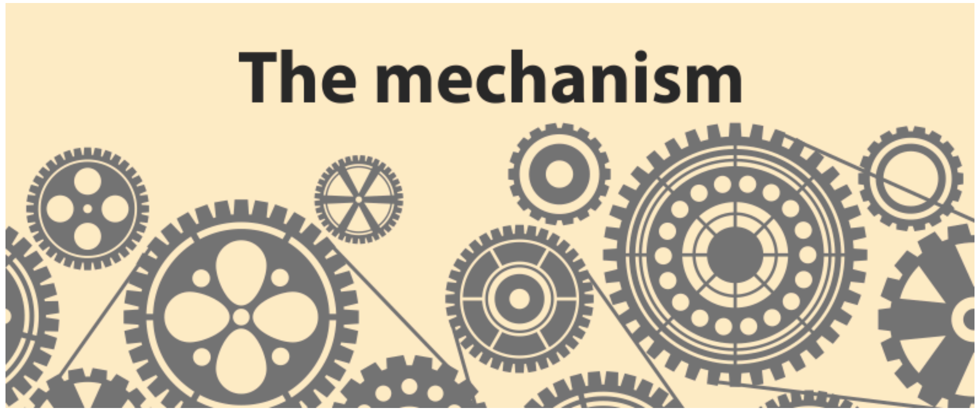
Nice Properties of Differential Privacy

- Privacy loss measure (ε)
  - Bounds the cumulative privacy losses across different computations and databases
- Resilience to arbitrary post-processing
  - Adversary's background knowledge is irrelevant
- Compositional reasoning
  - Programmability: construct complicated private analyses from simple private building blocks

Other Formulations

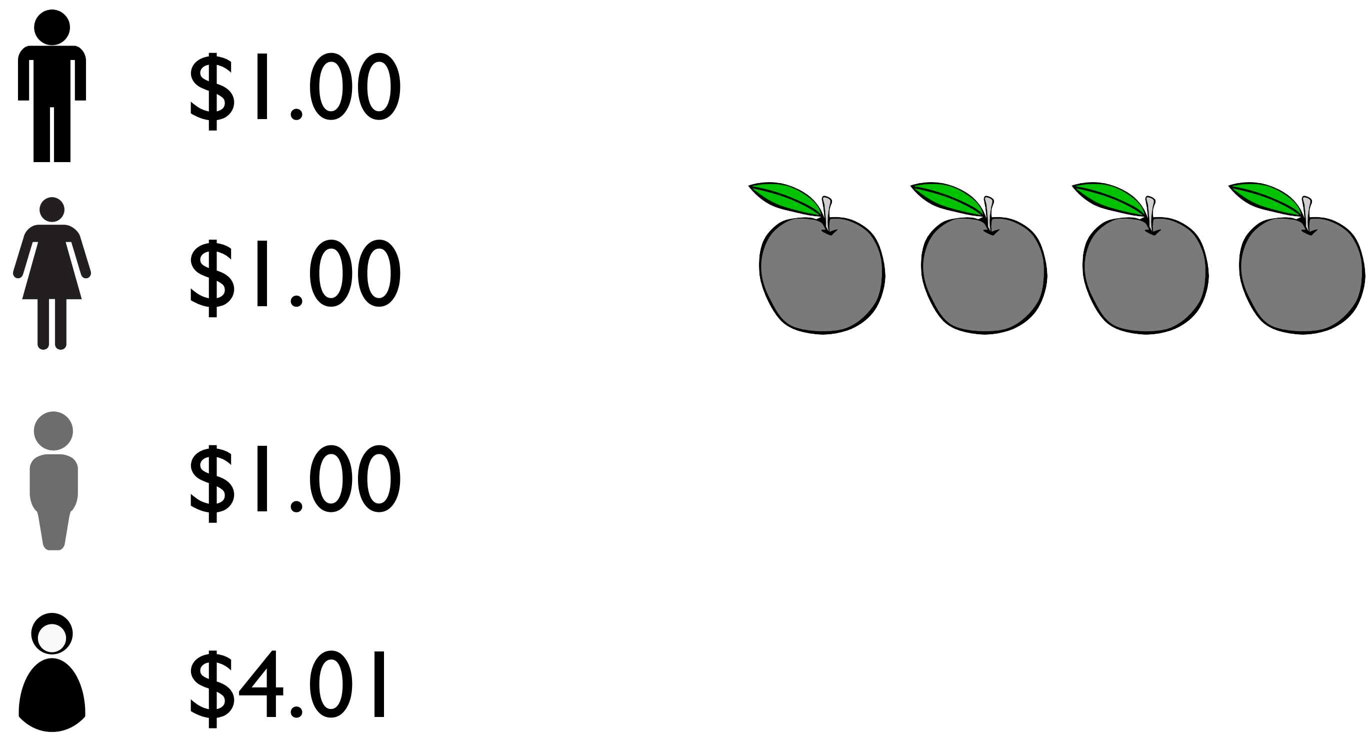
- Renyi Differential Privacy [Mir17]
- (Zero)-Concentrated Differential Privacy [DR16, BS16]
- Truncated-Concentrated Differential Privacy [BDRS18]

Privacy as a Tool for Mechanism Design



Warmup: Revenue Maximization

n buyers
w/ private value



- Could set the price of apples at $1.00 for profit: $4.00
- Could set the price of apples at $4.01 for profit $4.01

  - Best price: $4.01, 2nd best price: $1.00
  - Profit if you set the price at $4.02: $0
  - Profit if you set the price at $1.01: $1.01

Incentivizing Truth-telling

- A mechanism $M: \mathscr{X}^n \to \mathscr{R}$ for some abstract range $\mathscr{R}$
  - $\mathscr{X}$ = *reported* value; $\mathscr{R}$ = {$1.00, $1.01, $1.02, $1.03, …}
- Each agent i has a utility function $u_i: \mathscr{R} \to [-B, B]$
  - For example, $u_i(r) = \mathbf{1}[x \geq r](v - r)$, if r is the selected price

Definition. A mechanism M is α-approximately *dominant strategy truthful* if for any i with private value v_i , any reported value x_i from i and any reported values from everyone else x_{-i}

$$\mathbb{E}_M[u_i(M(v_i, x_{-i}))] \geq \mathbb{E}_M[u_i(M(x_i, x_{-i}))] - \alpha$$

No matter what other people do,
truthful report is (almost) the best

Privacy $\Rightarrow$ Truthfulness

- A mechanism $M\colon \mathscr{X}^n \to \mathscr{R}$ for some abstract range $\mathscr{R}$
- Each agent i has a utility function $u_i\colon \mathscr{R} \to [-B, B]$

Theorem [MT07]. Any ϵ-differentially private mechanism M is ϵB-approximately *dominant strategy truthful.*

Proof idea.

Utilitarian view of the DP definition: for all utility function u_i

$$\mathbb{E}_M[u_i(M(x_i, x_{-i}))] \geq \exp(\epsilon)\, \mathbb{E}_M[u_i(M(x_i', x_{-i}))]$$

The Exponential Mechanism

[MT*07*]

- A mechanism $M: \mathscr{X}^n \rightarrow \mathscr{R}$ for some abstract range $\mathscr{R}$
  - $\mathscr{X}$ = *reported* value; $\mathscr{R}$ = {\$1.00, \$1.01, \$1.02, \$1.03, …}
- Paired with a *quality score* $q: \mathscr{X}^n \times \mathscr{R} \rightarrow \mathbb{R}$.
  - $q(D, r)$ represents how good output r is for input data D, (e.g., revenue)
  - Sensitivity Δq: for all neighboring D and D', $r \in \mathscr{R}$

$$|q(D,r) - q(D',r)| \leq \Delta q$$

The Exponential Mechanism

[MT *07*]

- Input: data set D, range $\mathscr{R}$, quality score q, privacy parameter ϵ
- Select a random outcome r with probability proportional to

$$\mathbb{P}[r] \propto \exp\left(\frac{\epsilon\, q(D,r)}{2\Delta q}\right)$$

Idea: Make high quality outputs *exponentially* more likely at a rate that depends on the sensitivity of the quality Δq and the privacy parameter ϵ

The Exponential Mechanism

[MT *07*]

- Input: data set D, range $\mathscr{R}$, quality score q, privacy parameter ϵ
- Select a random outcome r with probability proportional to

$$\mathbb{P}[r] \propto \exp\left(\frac{\epsilon\, q(D, r)}{2\Delta q}\right)$$

Theorem [MT07]. The exponential mechanism is ϵ-differentially private, $O(\epsilon)$-approximately DS truthful and with probability $1 - \beta$, the selected outcome $\hat{r}$ satisfies

$$q(D, \hat{r}) \geq \text{OPT} - \frac{2\Delta q \, \log(|\mathscr{R}|/\beta)}{\epsilon}$$

Limitations

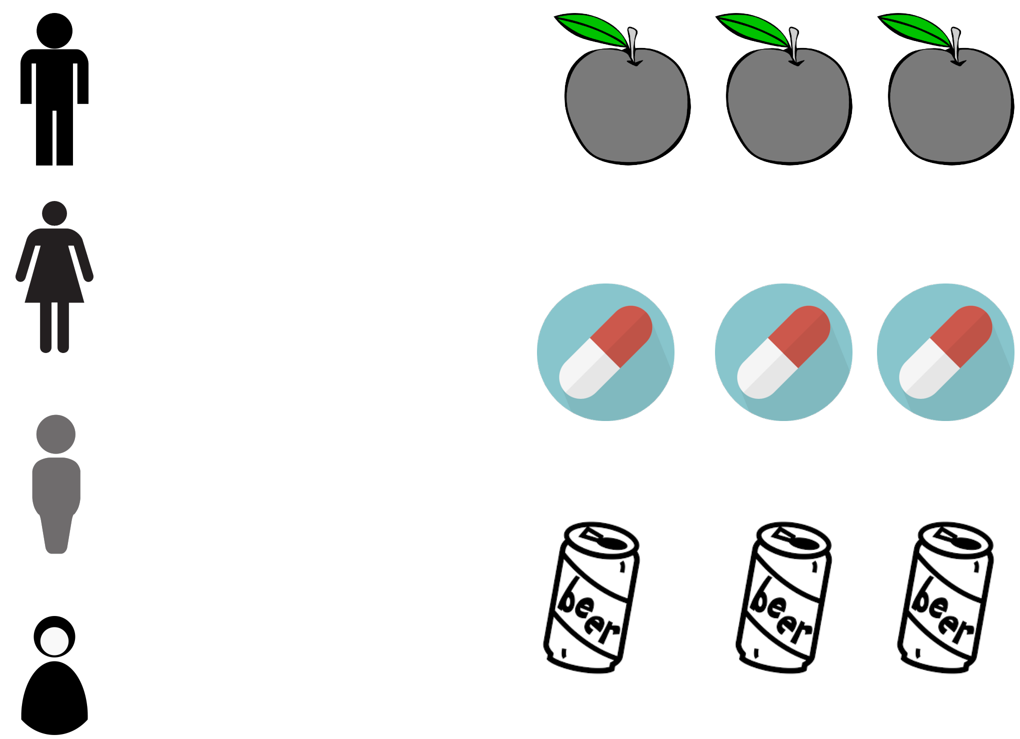
- *Everything* is an approximate dominant strategy, not just truth telling.
  - Sometimes it is easy to find a beneficial deviation
  - [NST12, HK12] obtain exact truthfulness

- Many interesting problems cannot be solved under the standard constraint of differential privacy

Joint Differential Privacy as a Tool

Allocation Problem

n buyers



k types of goods
s copies of each

Each buyer i has private value $v_i(j) = v_{ij}$ for each good j

Mechanism Design Goal

- Design a mechanism M that computes a feasible allocation $x_1, \ldots, x_n$ and a set of item prices $p_1, \ldots, p_k$ such that

- The allocation maximizes social welfare

$$SW = \sum_{i=1}^{n} v_i(x_i)$$

- α-approximately dominant strategy truthful

$$\mathbb{E}_{M(V')}[v_i(x_i) - p(x_i)] \leq \mathbb{E}_{M(V)}[v_i(x_i) - p(x_i)] + \alpha$$

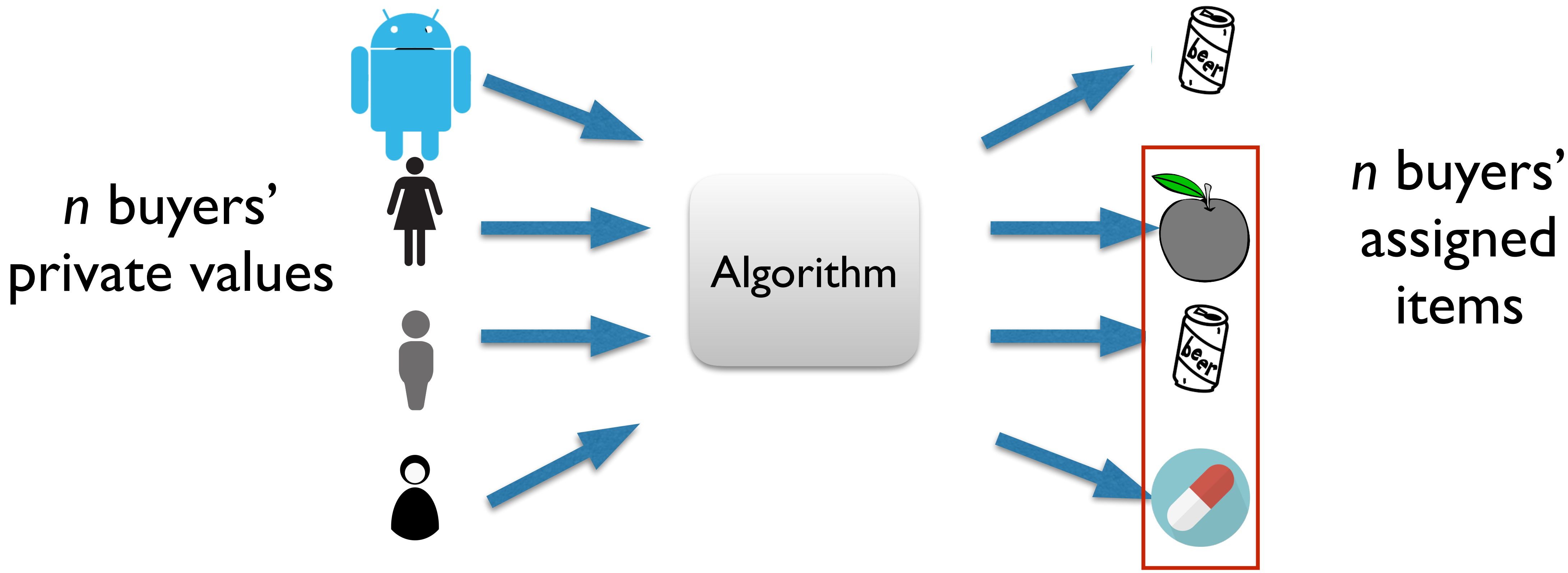
for any $V = (v_1, \ldots, v_i, \ldots, v_n)$ and $V' = (v_1, \ldots, v_i', \ldots, v_n)$

Using Privacy as a Hammer?

Impossible to solve under standard differential privacy

- Output of the algorithm: assignment of items to the buyers
- Differential privacy requires the output to be insensitive to change of any buyer's private valuation
- But to achieve high welfare, we will have to give the buyers what they want

Structure of the Problem

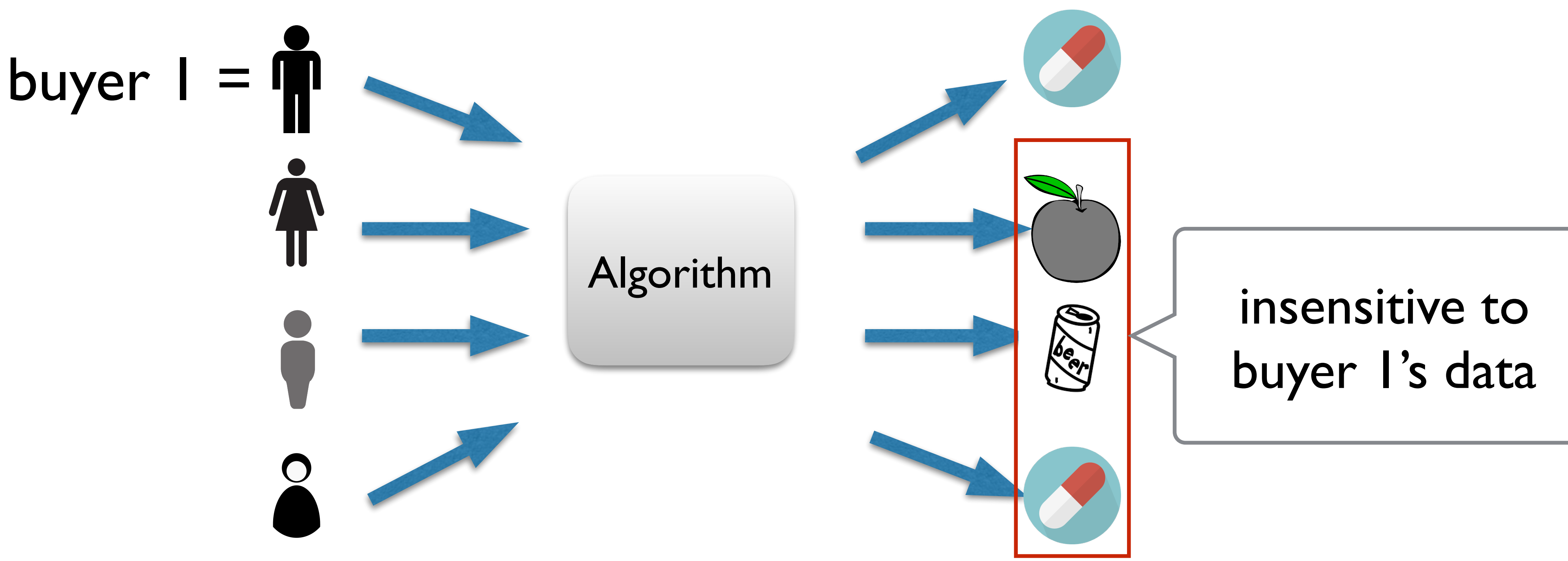


- Both the input and output are partitioned amongst *n* buyers
- The next best thing: protect a buyer's privacy from all other buyers

Joint Differential Privacy (JDP)
[KPRU14]

Definition: Two inputs D, D' are *i*-neighbors if they only differ by *i*'s input. An algorithm $A: X \to R^n$ satisfies (ε, δ)-joint differential privacy if for all neighbors D, D' and every $S \subseteq R^{n-1}$

$$\Pr[A(D)_{-i} \in S] \leq e^{\varepsilon} \Pr[A(D')_{-i} \in S] + \delta$$



Even if all the other buyers collude, they will not learn about buyer 1's private values!

How to solve the allocation problem under joint differential privacy?

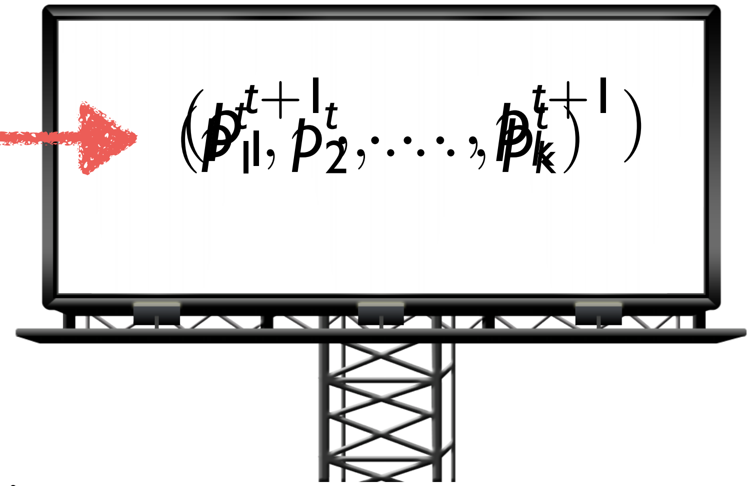
[HHRRW*14*, HHRW*16*]

Key idea:

use prices under *standard* differential privacy as a coordination device among the buyers

Price Coordination under JDP

"Billboard"



Price (Dual)

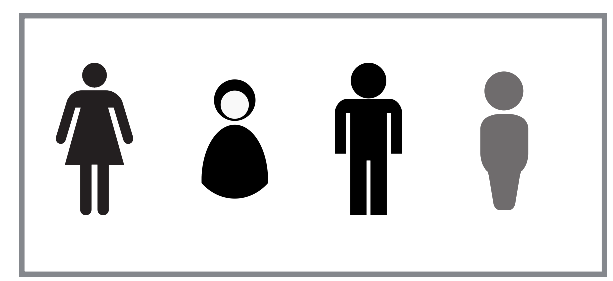
Iteratively updates prices

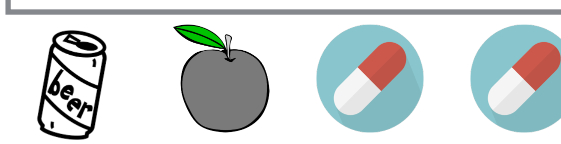
- Perturb the gradient (for privacy)
- Gradient descent update on the prices
- Raise prices on over-demand goods
- lower prices on under-demand goods

Buyers (Primal)

best response

Buyers best respond to prices *separately*





Demand the favorite item given the prices

The aggregate demand gives gradient feedback

Final Solution (average allocation): Let each buyer uniformly randomly sampled an item from the sequence of best responses

Approximate Truthfulness

Incentivize truth-telling with privacy

- Final prices are computed under differential privacy (insensitive to any single buyer’s misreporting)
- Each buyer is getting the (approximately) most preferred assignment given the final prices
- Truthfully reporting their data is an approximate dominant strategy for all buyers

Extension to Combinatorial Auctions

Allocating bundles of goods

- [HHRRW14] Gross substitutes valuations
- [HHRW16] d-demand valuations
  (general valuation over bundles of size at most d)

Compared to VCG mechanism

- JDP gives item prices; VCG charges payments on bundles
- JDP approximate envy-free; VCG not envy-free

Joint Differential Privacy as a Hammer

Meta-Theorem [KPRU14]

Computing equilibria subject to joint differential privacy robustly incentivizes truth telling.

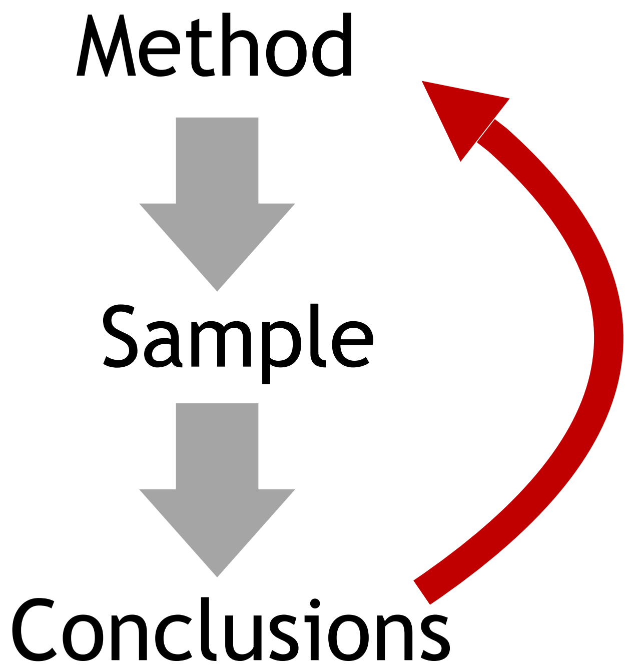
Solves *large-market* mechanism design problems for:

- [KMRW15] Many-to-one stable matching
    - First approximate *student-truthful* mechanism for approximate *school-optimal* stable matchings without distributional assumptions
- [RR14, RRUW15] Coordinate traffic routing (with tolls)
- [CKRW15] Equilibrium selection in anonymous games

Outline

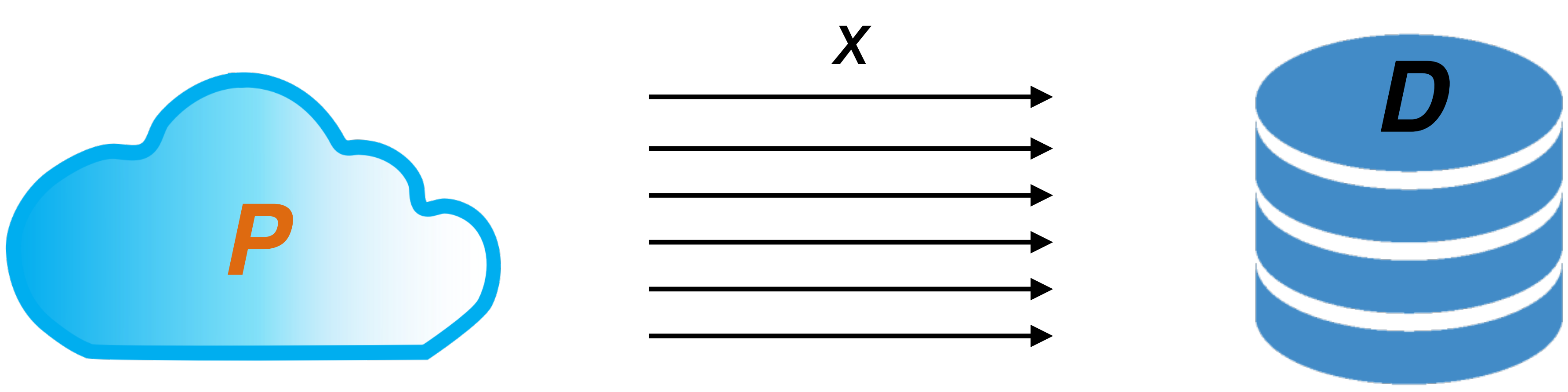
- Simple Introduction to Differential Privacy
- Mechanism Design
- Adaptive Data Analysis
- Certified Robustness

Adaptive Data Analysis



Basic Framework

- A data universe *X*
- A distribution *P* over *X*
- A dataset *D* consisting of *n* points *x* in *X* drawn i.i.d. from distribution *P*



n i.i.d. draws

Adaptivity in Learning

- Suppose we want to train a model to classify dogs and cats pictures…

A diligent data scientist

model 1 →

← error 0.4

model 2 →

← error 0.3

…

↓

Super refined model *M* with error 0.0001 on *D*

Data set drawn i.i.d. from *P*

Choosing a Formalism: Statistical Queries

- A *statistical query* is defined by a predicate

$$\phi : X \to [0,1]$$

- The value statistical query is

$$\phi(P) = \mathbb{E}_{x \sim P}[\phi(x)]$$

Generality

- Means, variances, correlations, etc.

- Risk of a hypothesis:

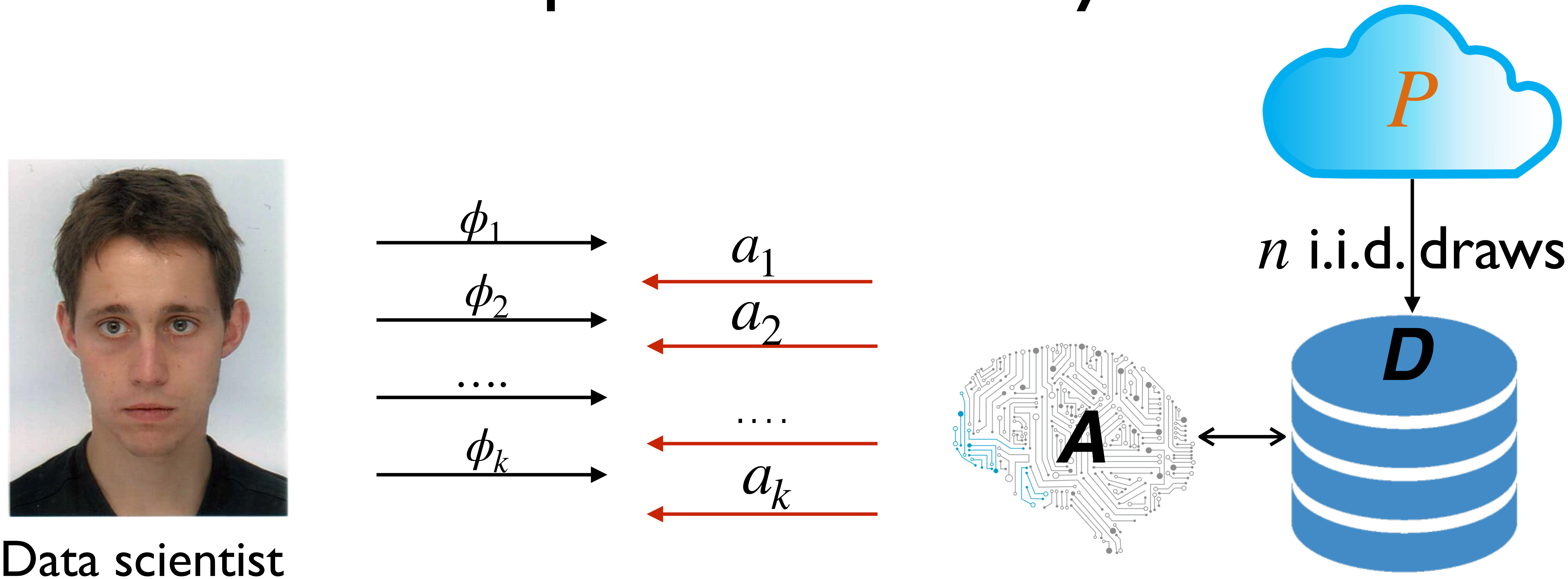
$$R(h) = \mathbb{E}_{(x,y)\sim P}[\ell(h(x), y)]$$

- *Gradient* of risk of a hypothesis:

$$\nabla R(h) = \mathbb{E}_{(x,y)\sim P}[\nabla \ell(h(x), y)]$$

- *Almost* all of PAC learning algorithms

Adaptive Data Analysis



Data scientist

Goal: Design A such that for all j

$$|a_j - \phi_j(P)| \leq \alpha$$

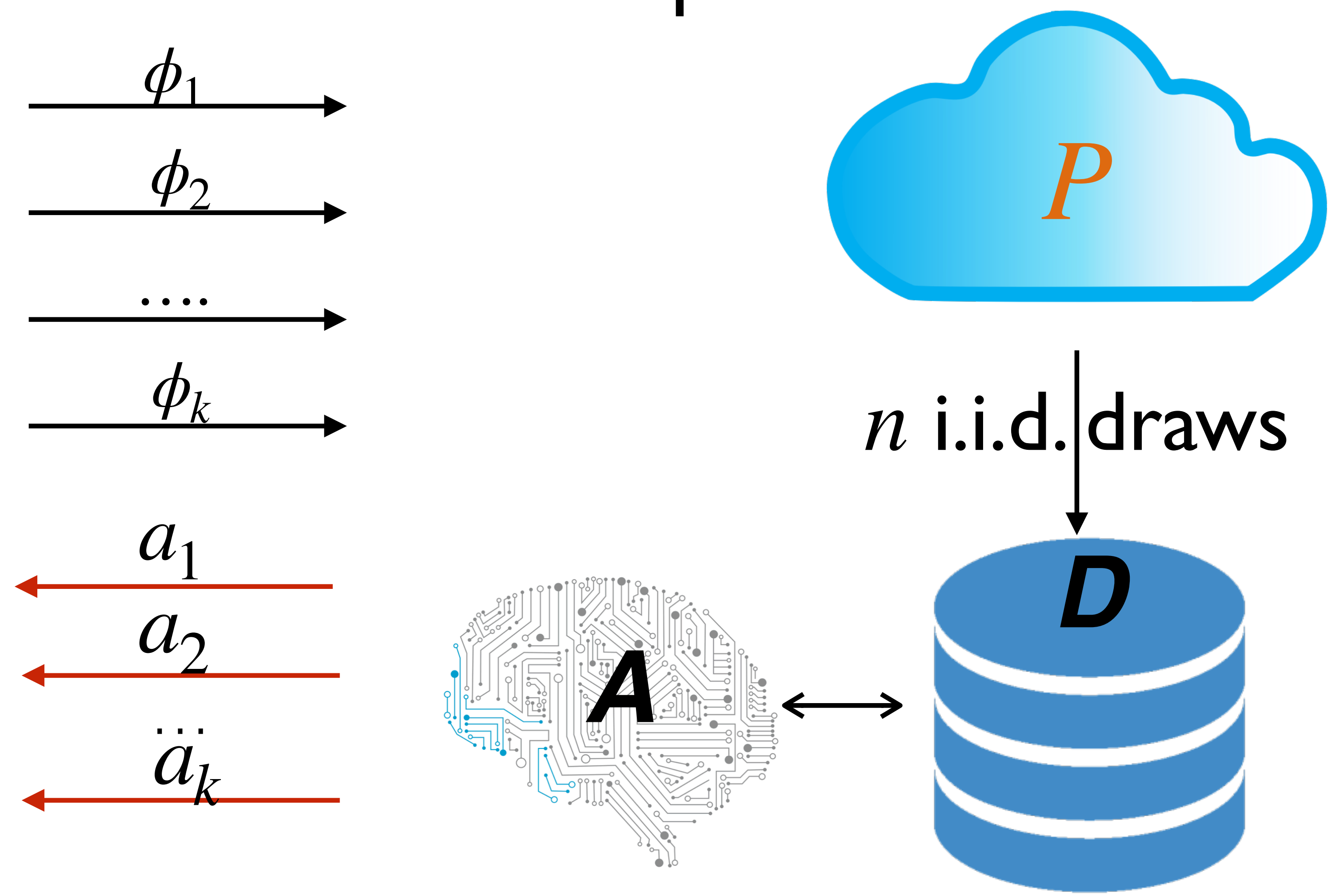
Challenge:

- A does not observe P
- Each ϕ_j depends arbitrarily on $q_1, a_1, \ldots, \phi_{j-1}, a_{j-1}$

Non-Adaptive Baseline

- Suppose the queries are chosen up front.

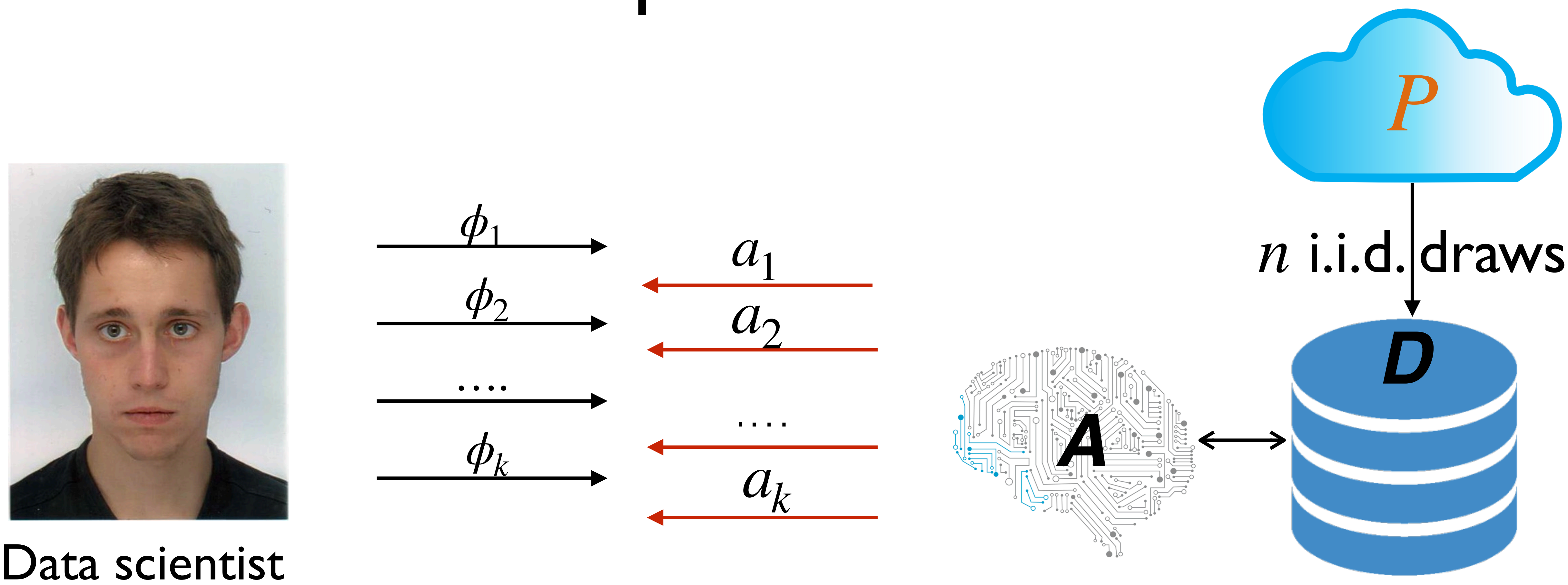
A well-behaved data scientist



The "empirical average" mechanism: $A_D(\phi) = \phi(D) = \frac{1}{n}\sum_{x \in D} \phi(x)$

$$\max_j |A_D(\phi_j) - \phi_j(P)| \lesssim \frac{\sqrt{\log k}}{\sqrt{n}}$$

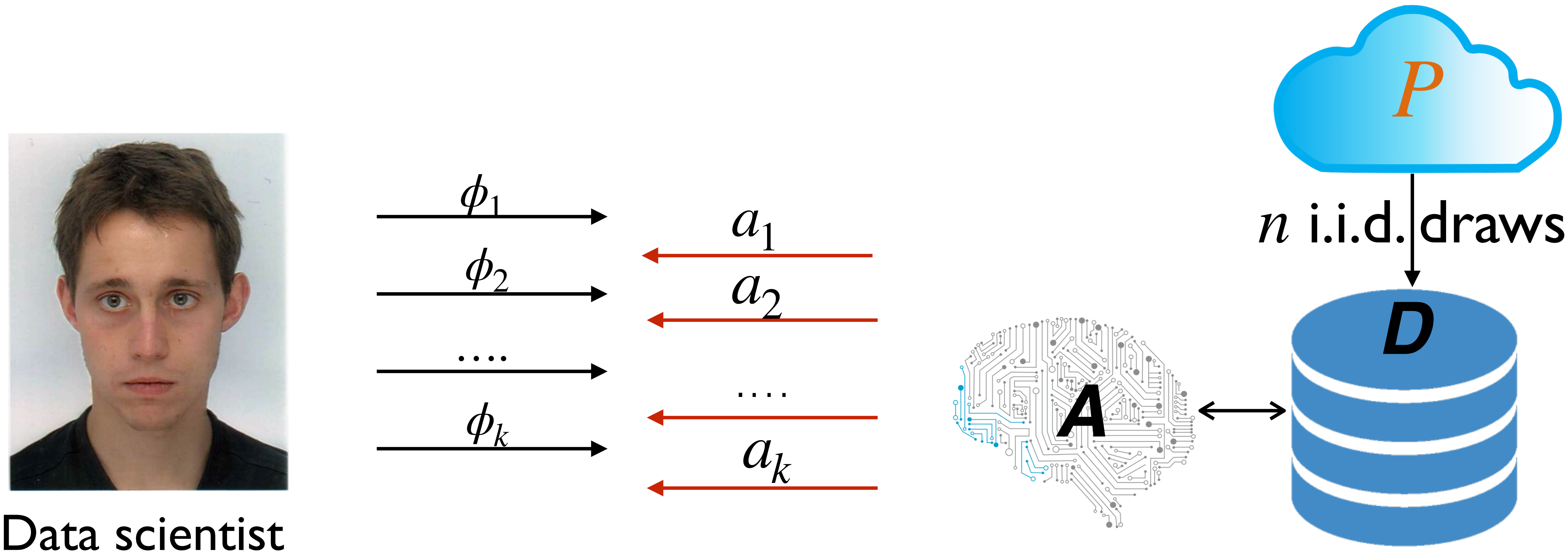
Adaptive Baseline



The "empirical average" mechanism: $A_D(\phi) = \phi(D) = \frac{1}{n}\sum_{x \in D} \phi(x)$

$$\max_j |A_D(\phi_j) - \phi_j(P)| \lesssim \frac{\sqrt{k}}{\sqrt{n}}$$

Improvement with Differential Privacy



The "noisy empirical" mechanism: $A_D(\phi) = \phi(D) + N(0,\sigma^2)$

$$\max_j |A_D(\phi_j) - \phi_j(P)| \lesssim \frac{k^{1/4}}{\sqrt{n}}$$

Adding noise reduces the error!

Gaussian Mechanism

Theorem [DFHPRR15, BNSSSU16, JLNRSS20]

The Gaussian mechanism can answer k adaptive SQs with error

$$\alpha = \tilde{O}\left(\frac{k^{1/4}}{\sqrt{n}}\right)$$

$$\frac{\sqrt{k}}{n\sigma} + \sigma$$

Can extend to other types of queries

- Lipchitz queries: $|q(D) - q(D')| \leq 1/n$
- Minimization queries: $q(D) = \arg\min_{\theta \in \Theta} \ell(\theta; D)$
- Bounded variance queries [FS17,18]

Proof sketch

[JLNRSS20]

- Data set $D \sim P^n$
- π : transcript between algorithm and analyst
  (sequence of query-answer pairs: $\phi_1, a_1, \ldots, \phi_k, a_k$)
- $Q_\pi = (P^n) \mid \pi$: "posterior" distribution conditioned on π
- Resample a new data set $S \sim Q_\pi$

Resampling Lemma

(D, π) and (S, π) are identically distributed

- π : transcript $(\phi_1, a_1, \ldots, \phi_k, a_k)$
- $Q_\pi = (P^n) \mid \pi$: "posterior" distribution conditioned on π
- Resample a new data set $S \sim Q_\pi$

Resampling Lemma

(D, π) and (S, π) are identically distributed

- A promises sample accuracy w.h.p. $|a_i - \phi_i(D)|$ is small
- By Resampling Lemma, $|a_i - \phi_i(Q_\pi)|$ is small

  where $\phi_i(Q_\pi) = \mathbb{E}_{S \sim Q_\pi}[\phi_i(S)]$

Now we know $|a_i - \phi_i(Q_\pi)|$ is small

where $\phi_i(Q_\pi) = \mathbb{E}_{S \sim Q_\pi}[\phi_i(S)]$

If the transcript π satisfies ϵ-differential privacy, then for any ϕ

$$\phi(Q_\pi) \leq e^{\epsilon}\, \phi(P)$$

$$\Rightarrow |\phi(Q_\pi) - \phi(P)| \leq e^{\epsilon} - 1 \approx \epsilon$$

Stronger Bounds

Theorem [DFHPRR15, BNSSSU16, JLNRSS20]

There exists a mechanism can answer k adaptive SQs with error

$$\alpha = \tilde{O}\left(\min\left\{\frac{k^{1/4}}{\sqrt{n}}, \frac{d^{1/6}\sqrt{\log k}}{n^{1/3}}\right\}\right)$$

- Dependence on d: data dimensionality
  - Unavoidable dependence [HU14, SU15]
- Uses a more powerful algorithm, namely PrivateMW [HR10]
- Computational issue: exponential in d

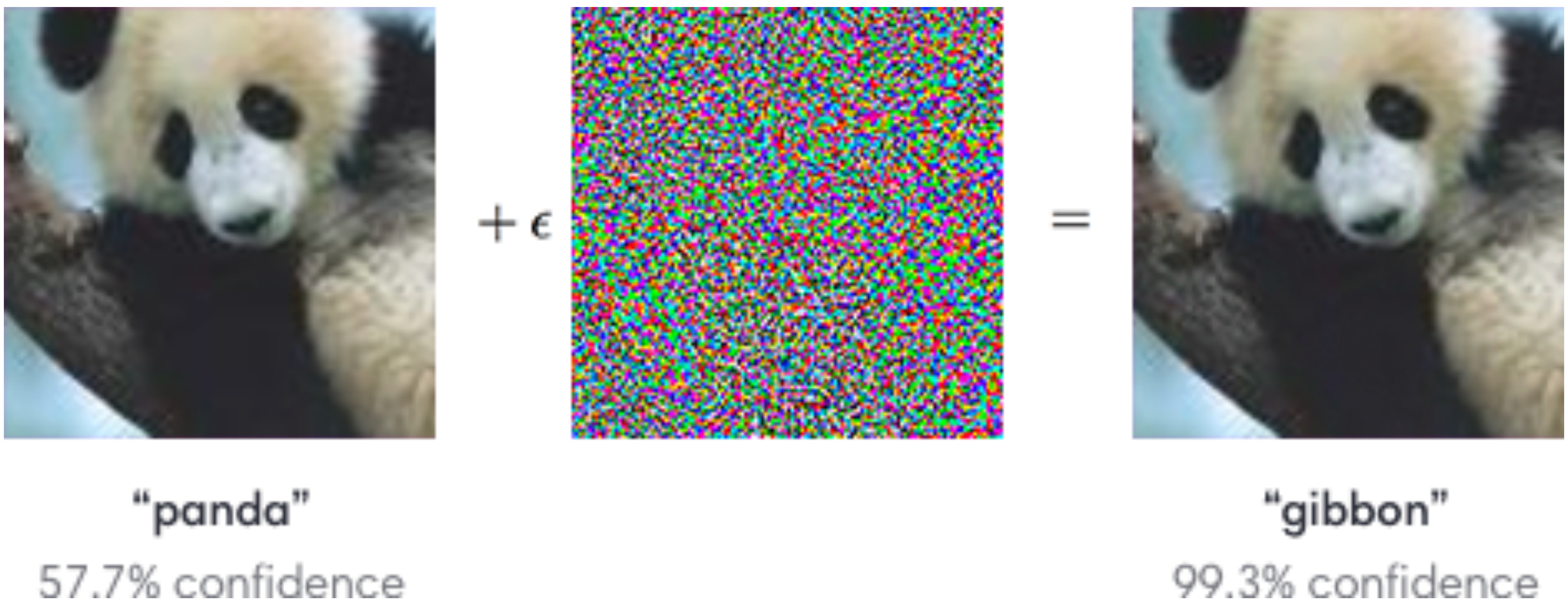
Other Applications

- Algorithmic application: Improve sample complexity
  - [HKRR18]: Enforcing Multi-calibration as fairness criterion
- Prove concentration inequalities [SU17,NS17]

Outline

- Simple Introduction to Differential Privacy
- Mechanism Design
- Adaptive Data Analysis
- Certified Robustness

Connection with Certified Robustness



[Goodfellow et al. 15]

Adversarial Example

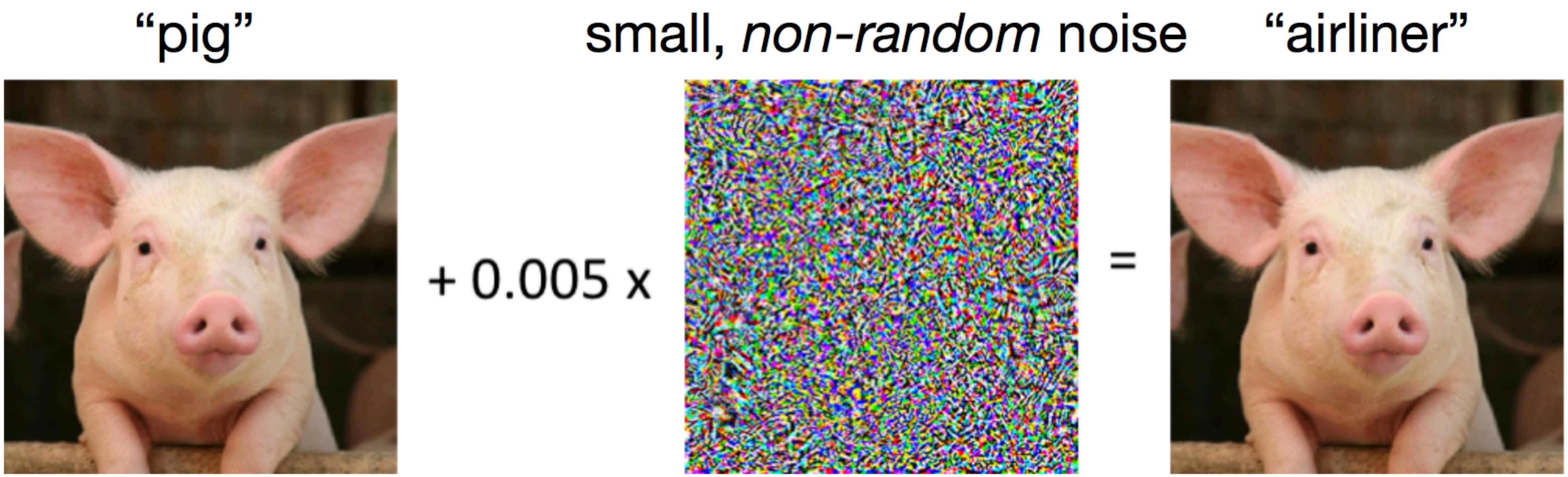


Figure from [Mądry et al.18]

Formulation

- (Hard) classifier $f\colon \mathbb{R}^d \to Y$
- *Soft* classifier $g: \mathbb{R}^d \to \Delta(Y)$
- Perturbation set S (e.g., ℓ_p ball of radius r)

A classifier g is robust to perturbations in S at example $x \in \mathbb{R}^d$ if

$$\arg\max_{c \in Y} g(x)_c = \arg\max_{c \in Y} g(x+\delta)_c \quad \text{for all} \quad \delta \in S$$

For this talk, $S = B_2(r)$.

Would like to tolerate large r

Two Approaches

- Empirical defenses
  - Adversarial training and variants
  - Performs well in practice, but no provable guarantees
- Certified robustness
  - Provable guarantees, but tend to perform worse in practice

PixelDP

[Lecuyer et al. 2018]

- Perturb each example x with Gaussian noise $\eta \sim N(0, \sigma^2 I)$
- Evaluate the prediction with the base classifier $f(x + \eta)$
- The prediction is differentially private in the pixels

For any x and x' such that $\|x - x'\|_2 \leq r$ and any $E \subseteq Y$

$$\mathbb{P}[f(x + \eta) \in E] \leq e^{\epsilon}\, \mathbb{P}[f(x' + \eta) \in E] + \delta$$

Even if $f(x) \neq f(x')$, the distributions satisfy

$$f(x + \eta) \approx f(x' + \eta)$$

Randomized Smoothing

Smoothed Classifier

$$g(x)_c = \mathbb{P}_{\eta \sim N(0, \sigma^2 I)}[f(x + \eta) = c]$$

Certified Robustness [Lecuyer et al. 18]

For any example $x \in \mathbb{R}^d$, if there exists a class c such that

$$g(x)_c > e^{2\epsilon} \max_{y \neq c} g(x)_y + (1 + e^{\epsilon})\,\delta$$

Then g is robust at x for any ℓ_2 perturbation of size

$$r \leq \frac{\sigma\epsilon}{\sqrt{2 \log(1.25/\delta)}}$$

Improved Bounds

Subsequently improved by [Li et al. 18] and [Cohen et al.19]

Theorem [Cohen et al. 19]

Fix any example $x \in \mathbb{R}^d$. Let g be the smoothed classifier of f. Let

$$a = \arg\max_{c \in Y} g(x)_c, \quad p_a = g(x)_a$$

$$b = \arg\max_{c \in Y, c \neq a} g(x)_c, \quad p_b = g(x)_b$$

Then g is robust at x for any ℓ_2 perturbation of size

$$r = \frac{\sigma}{2}\left(\Phi^{-1}(p_a) - \Phi^{-1}(p_b)\right)$$

Φ denotes the CDF of the standard Gaussian.

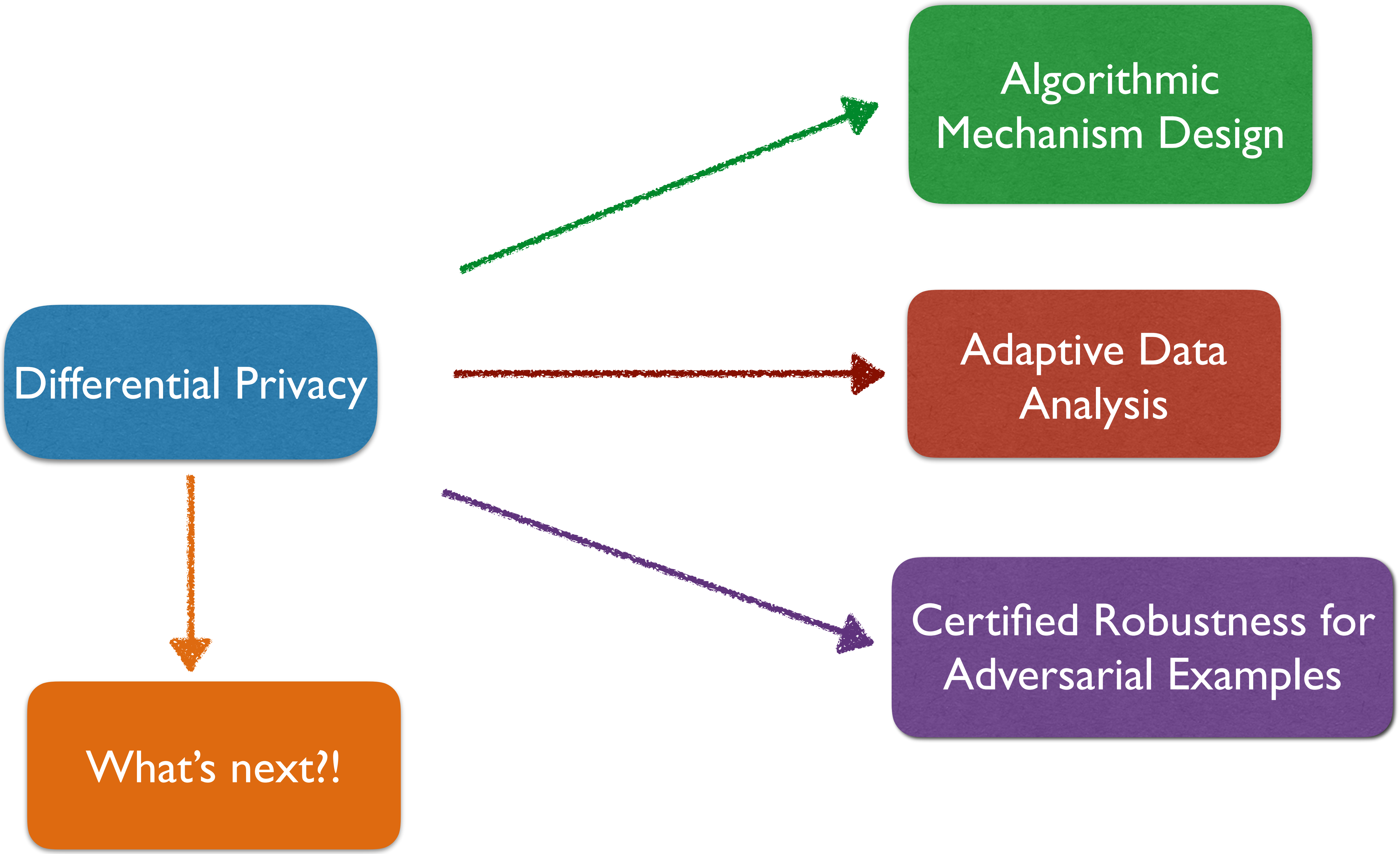
Proof using Neyman-Pearson lemma [NP33]

How about training?

[Salman et al.19]

- Beautiful idea of combining adversarial training with randomized smoothing

- Achieved SOTA certified accuracy for ℓ_2 perturbation


Algorithmic Mechanism Design
Differential Privacy
Adaptive Data Analysis
Certified Robustness for Adversarial Examples
What's next?!

Differential Privacy Techniques Beyond Differential Privacy

Steven Wu
Assistant Professor
University of Minnesota

Thanks Jerry Li, Aaron Roth and Jon Ullman
for their help with my slides!