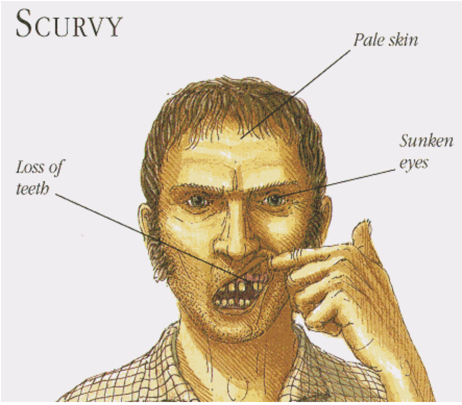
Order-Revealing Encryption and the Hardness of Private Learning

January 11, 2016

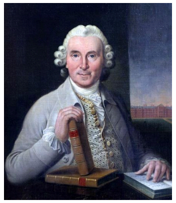
Mark Bun Harvard
Mark Zhandry MIT

Let's do some science!

- Scurvy: a problem throughout human history

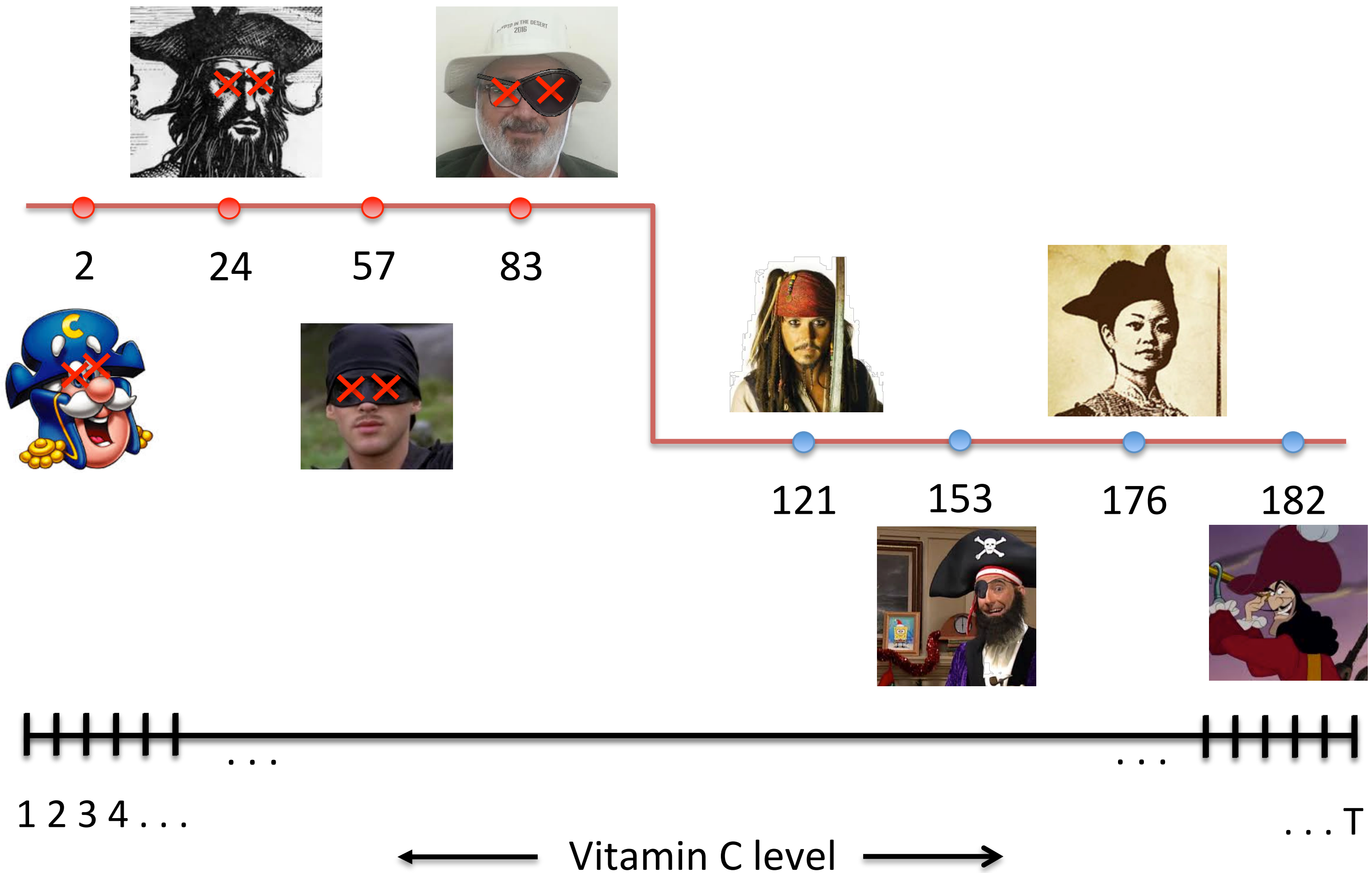


- Caused by vitamin C deficiency



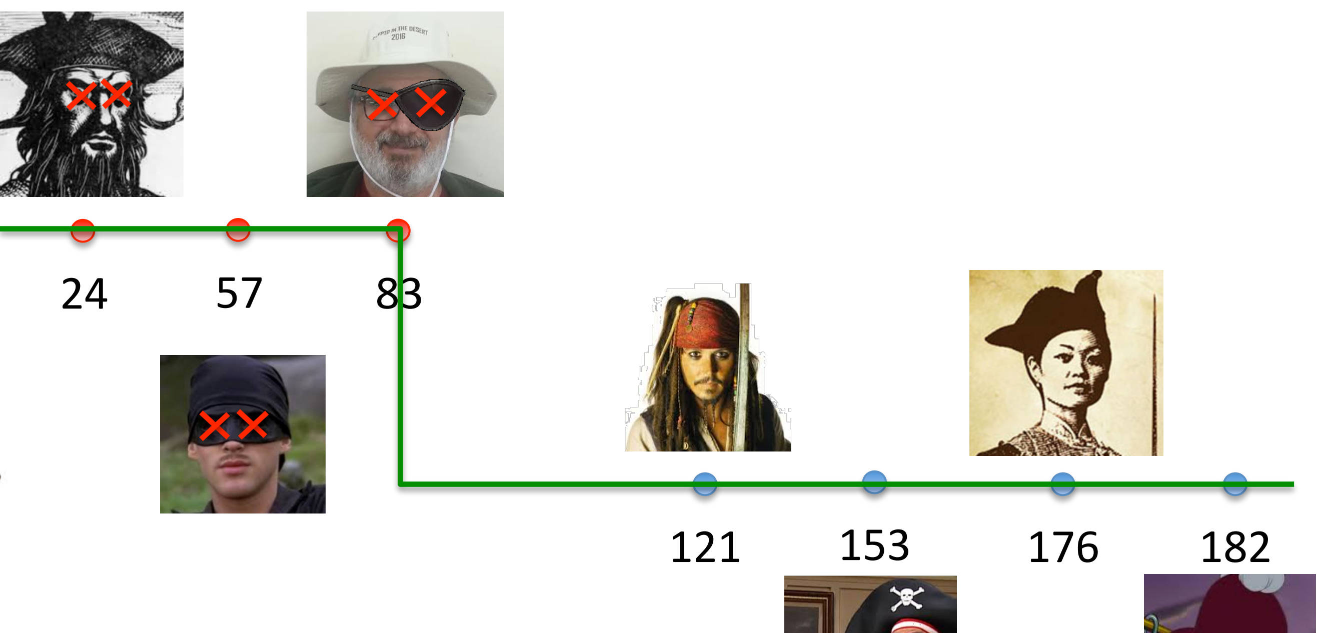
- How much vitamin C is enough?

So you collect some data...



So you collect some data…

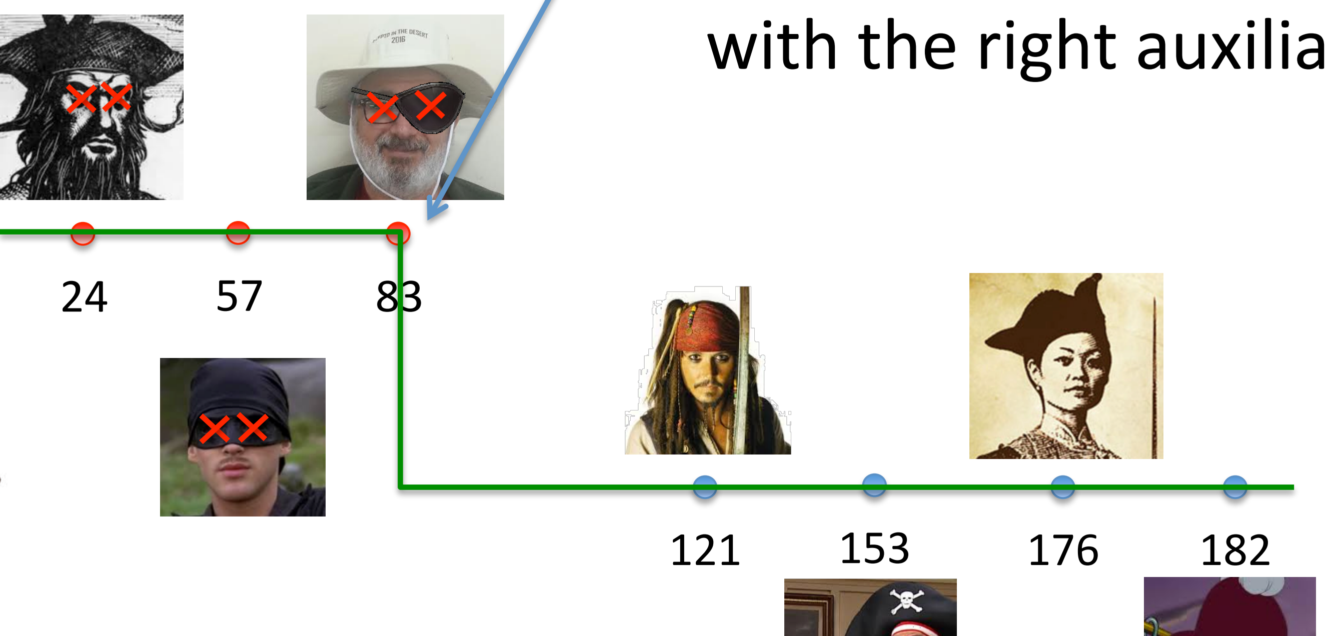
- Works for any #samples $n > n_0$
- Works for any threshold, on any underlying distribution



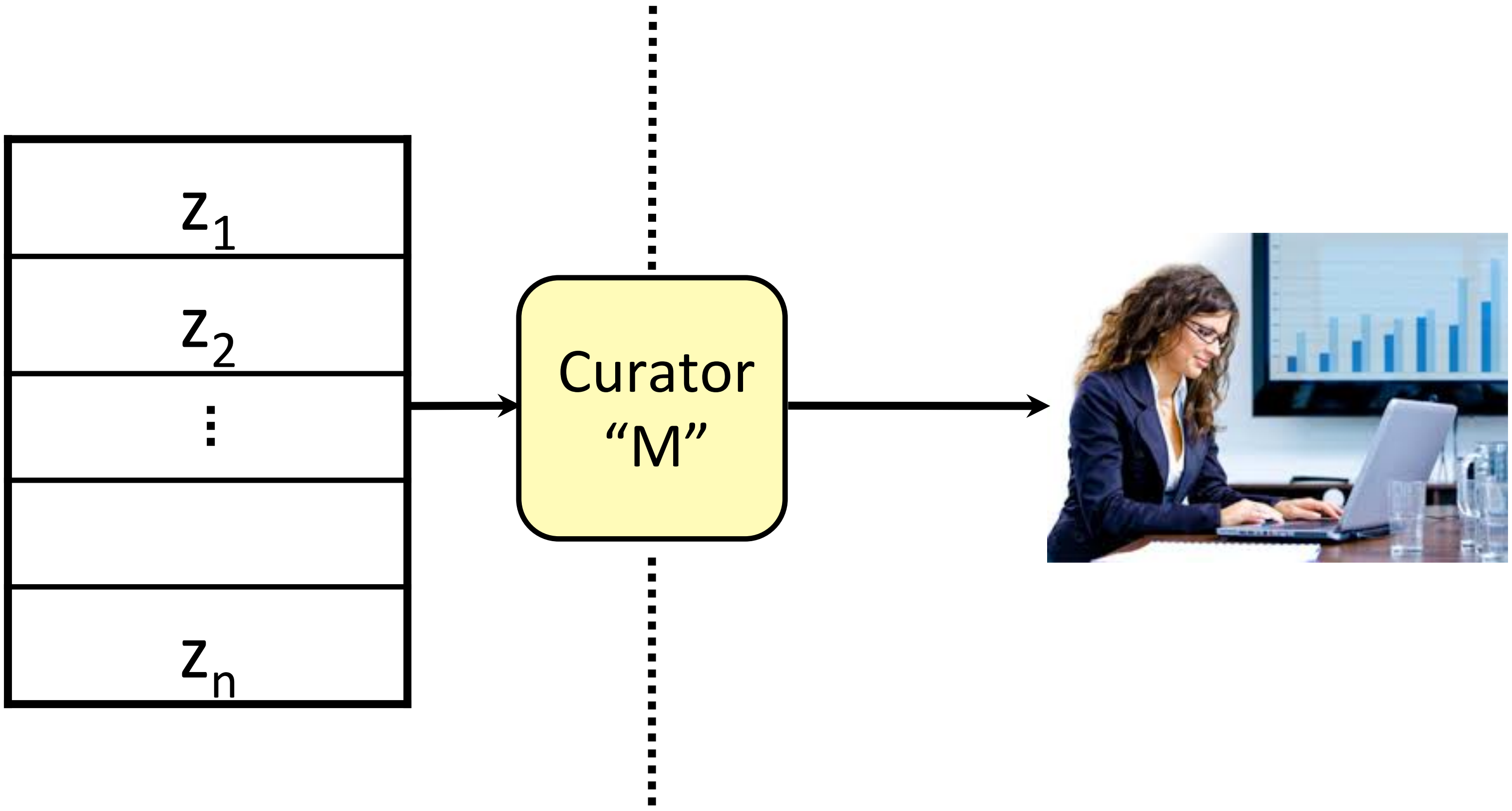
What’s the problem?

- The hypothesis threshold reveals someone’s data point!

- Could even be linked back to Kobbi with the right auxiliary info

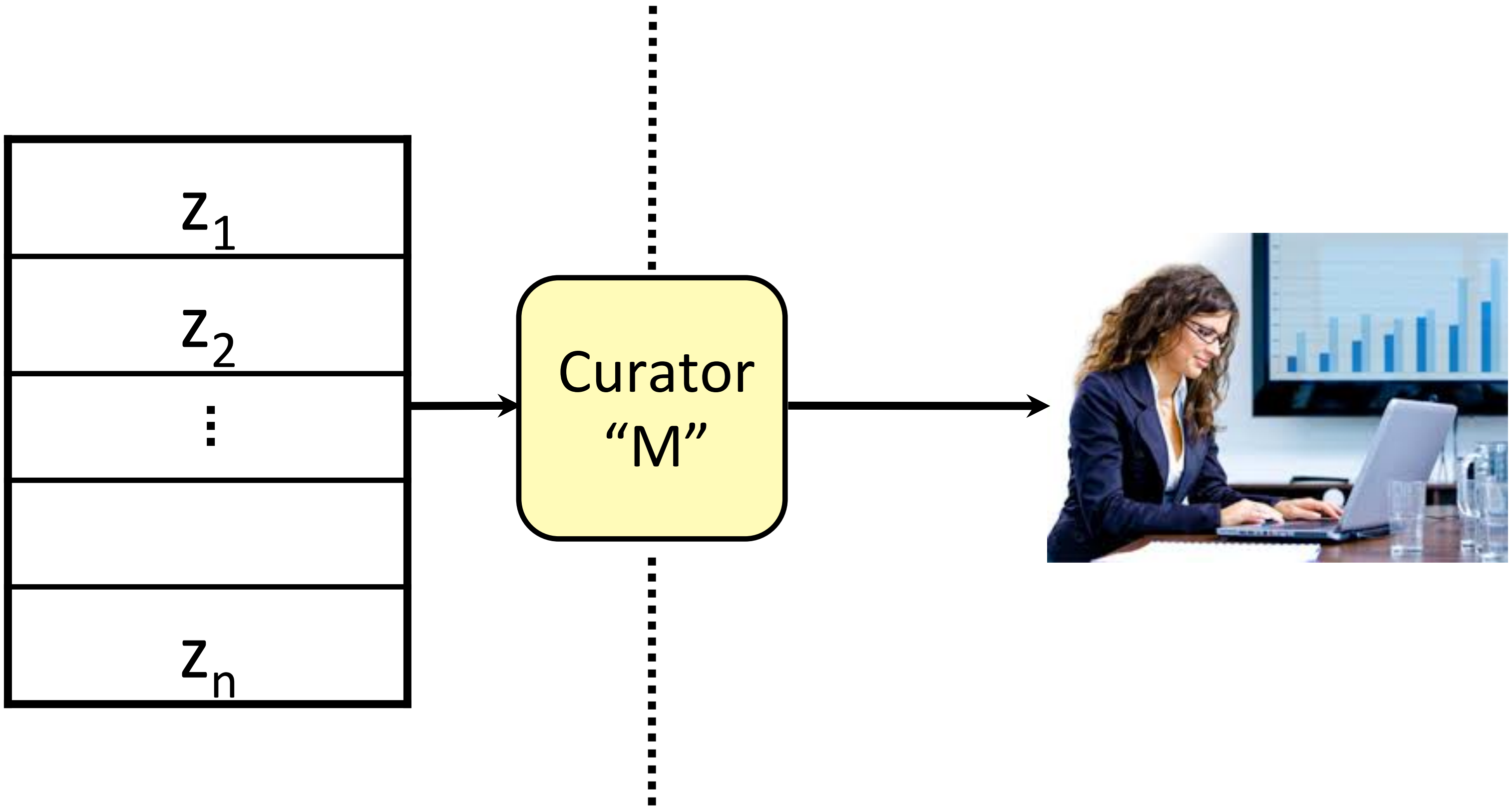


Privacy-Preserving Data Analysis



Want curators that are: ◆Private ◆Accurate ◆Efficient

Privacy-Preserving Data Analysis

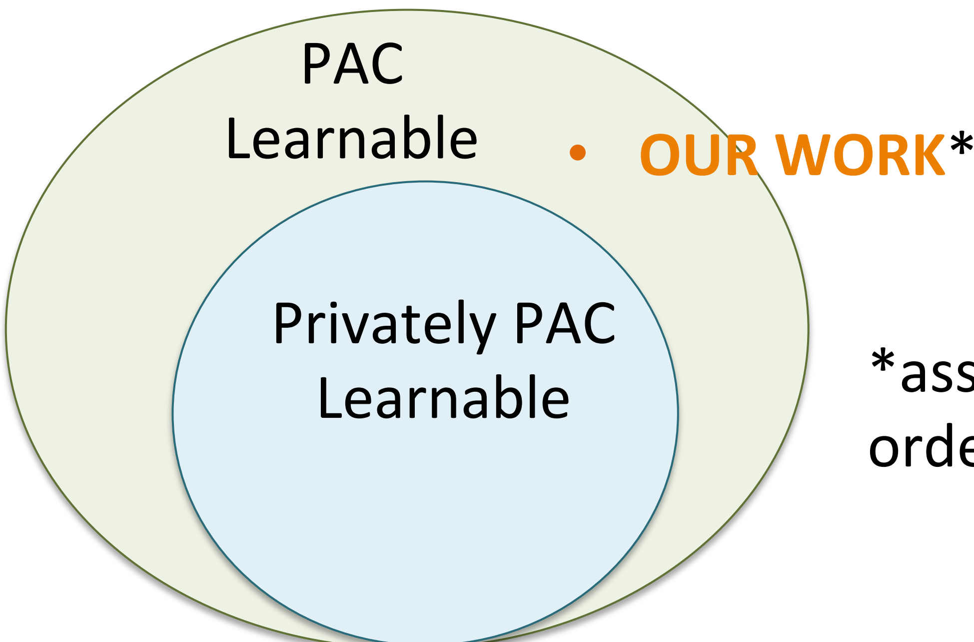


Want curators that are:

- Differentially Private
- Accurate Classifiers
- Computationally Efficient

This Talk

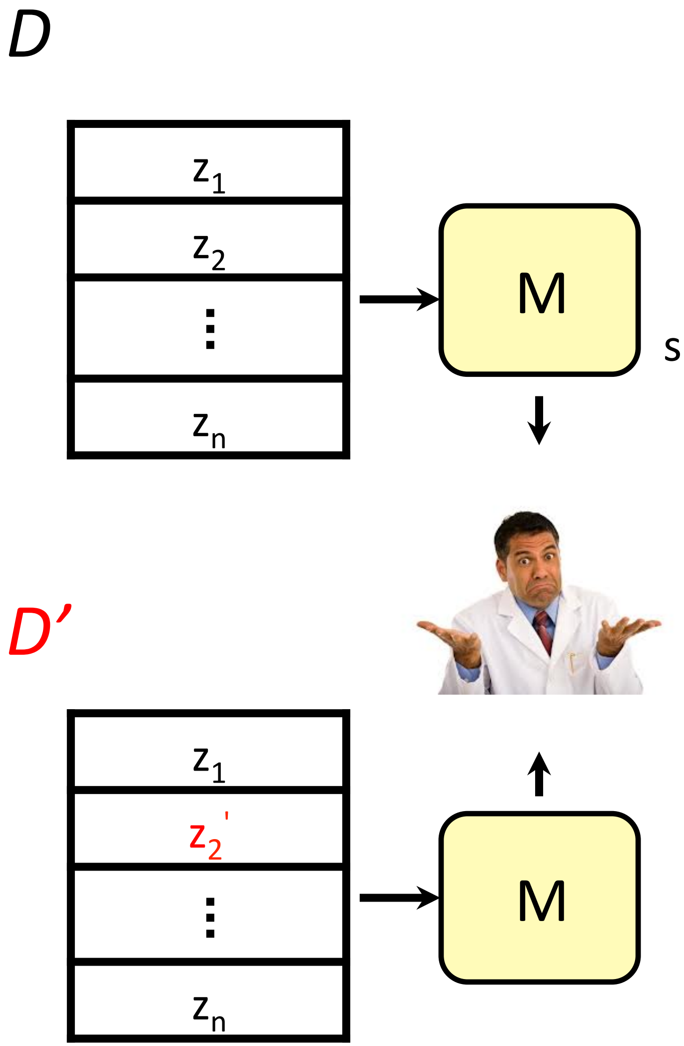
Computational complexity: Does private learning require more computational resources than non-private learning?



*assuming "strongly correct order-revealing encryption"

Differential Privacy

[Dinur-Nissim03+Dwork, Dwork-Nissim04, Blum-Dwork-McSherry-Nissim05, Dwork06, **Dwork-McSherry-Nissim-Smith06**, **Dwork-Kenthapadi-McSherry-Mironov-Naor06**]



D and *D′* are **neighbors** if they differ on one row

small const., e.g. $\varepsilon = 0.1$

$e^{\varepsilon} \approx 1 + \varepsilon$

"cryptographically small"

require $\delta \ll 1/n$, often $\delta = \text{negl}(n)$

M is **(ε,δ)-differentially private** if for all neighbors *D*, *D′* and S⊆Range(M):

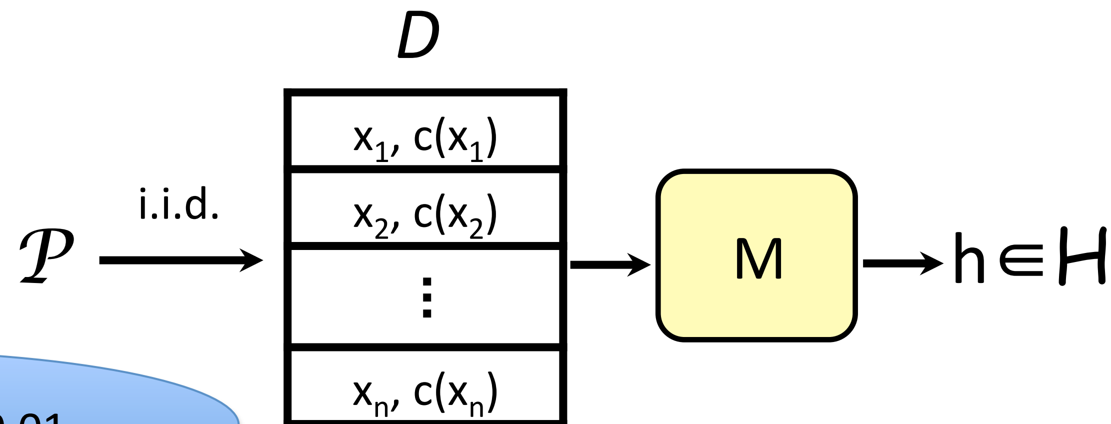
$$\Pr[M(D') \in S] \le e^{\varepsilon} \Pr[M(D) \in S] + \delta$$

◆Privacy ◆Accuracy ◆Complexity

PAC Learning [Valiant84]

$\mathcal{P}$ = unknown distribution over domain X

$\mathcal{C}$ = concept class {c: X→{0, 1}} $\mathcal{H}$ = hypothesis class {h: X→{0, 1}}



This talk: $\alpha = \beta = 0.01$

Hypothesis h is **α-good** if $\Pr_{x\sim\mathcal{P}}[h(x) \neq c(x)] \leq \alpha$

M is **(α,β)-accurate** if for all $\mathcal{P}$ and c, $\Pr_{M,D}[M(D) \text{ is } \alpha\text{-good}] \geq 1-\beta$

M is **efficient** if it runs in time poly(log|$\mathcal{C}$|, 1/α, 1/β)

◆Privacy ◆Accuracy ◆Complexity

Private PAC Learning

[Kasiviswanathan-Lee-Nissim-Raskhodnikova-Smith08]

(α, β)-PAC Learning

+ (ε, δ)-Differential Privacy

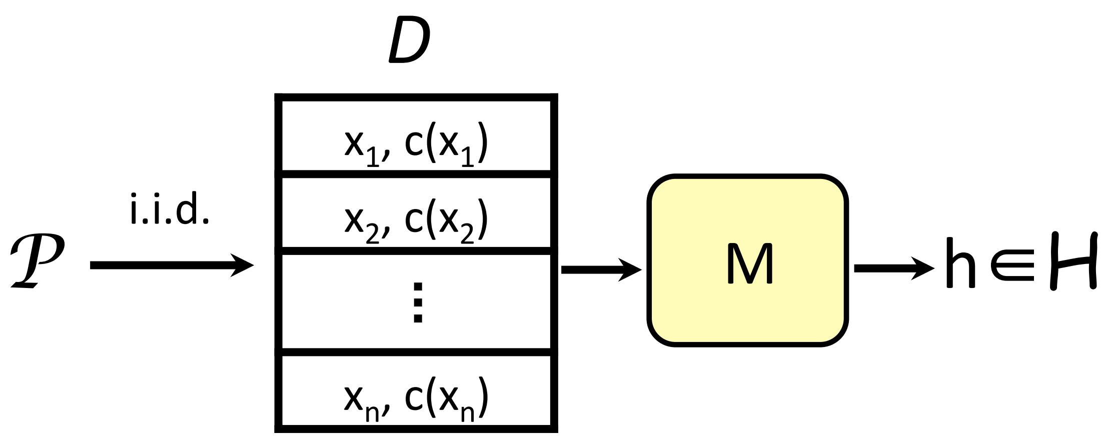
$(\alpha, \beta, \varepsilon, \delta)$-Private Learning

Private PAC Learning

[Kasiviswanathan-Lee-Nissim-Raskhodnikova-Smith08]

Algorithm M is a private learner if:

- M is an (α, β)-PAC learner for C
- M is (ε, δ)-differentially private



Slides thanks to Kobbi Nissim

◆Privacy ◆Accuracy ◆Complexity

Why Private Learning?

- Abstracts many statistical tasks that are performed on sensitive data

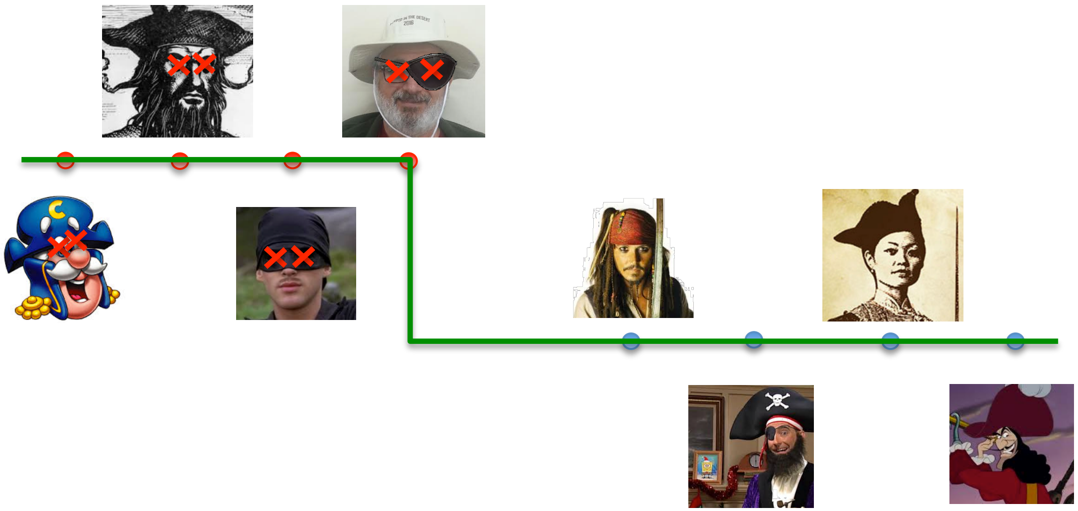
- Learning is intimately connected to privacy
  - Learning algorithms ⇒ DP algorithms [BLR08, HT10, HRS12]
  - Privacy ⇒ generalization [McSherry, DFHPRR15, BH15, BNSSSU15]

Slides thanks to Kobbi Nissim

What can be Learned Privately?

"Private Occam's Razor" [McSherry-Talwar07, KLNRS08]

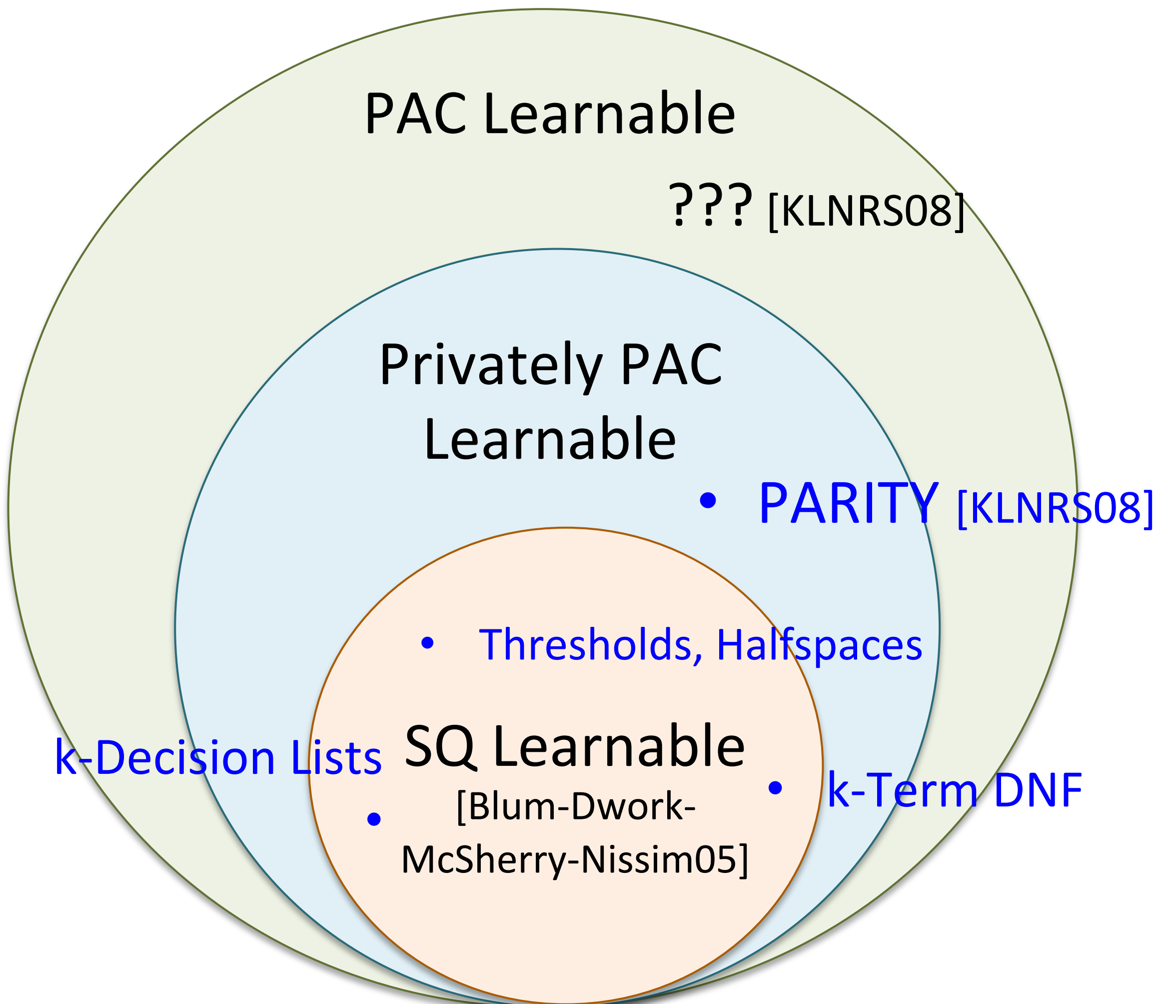
- Sample a nearly consistent hypothesis at random



- Thm: Any finite concept class C is privately learnable…
- …but in general, sampling is computationally inefficient

◆Privacy ◆Accuracy ◆Complexity

What can be Learned Privately *and* Efficiently?

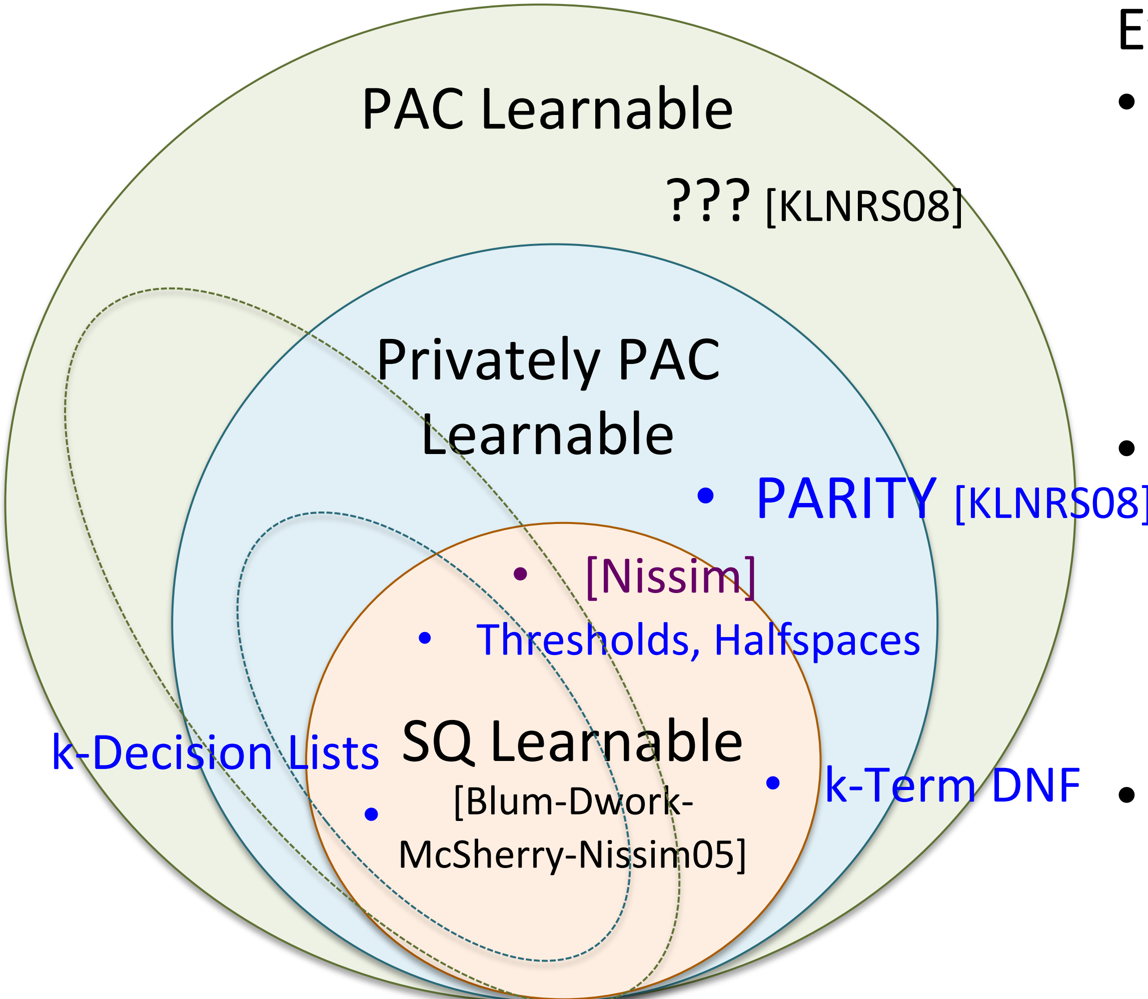


Known techniques for (efficient) PAC learning:

- Statistical Queries [Kearns93]
- Gaussian elimination for PARITY

◆Privacy ◆Accuracy ◆Complexity

What can be Learned Privately *and* Efficiently?

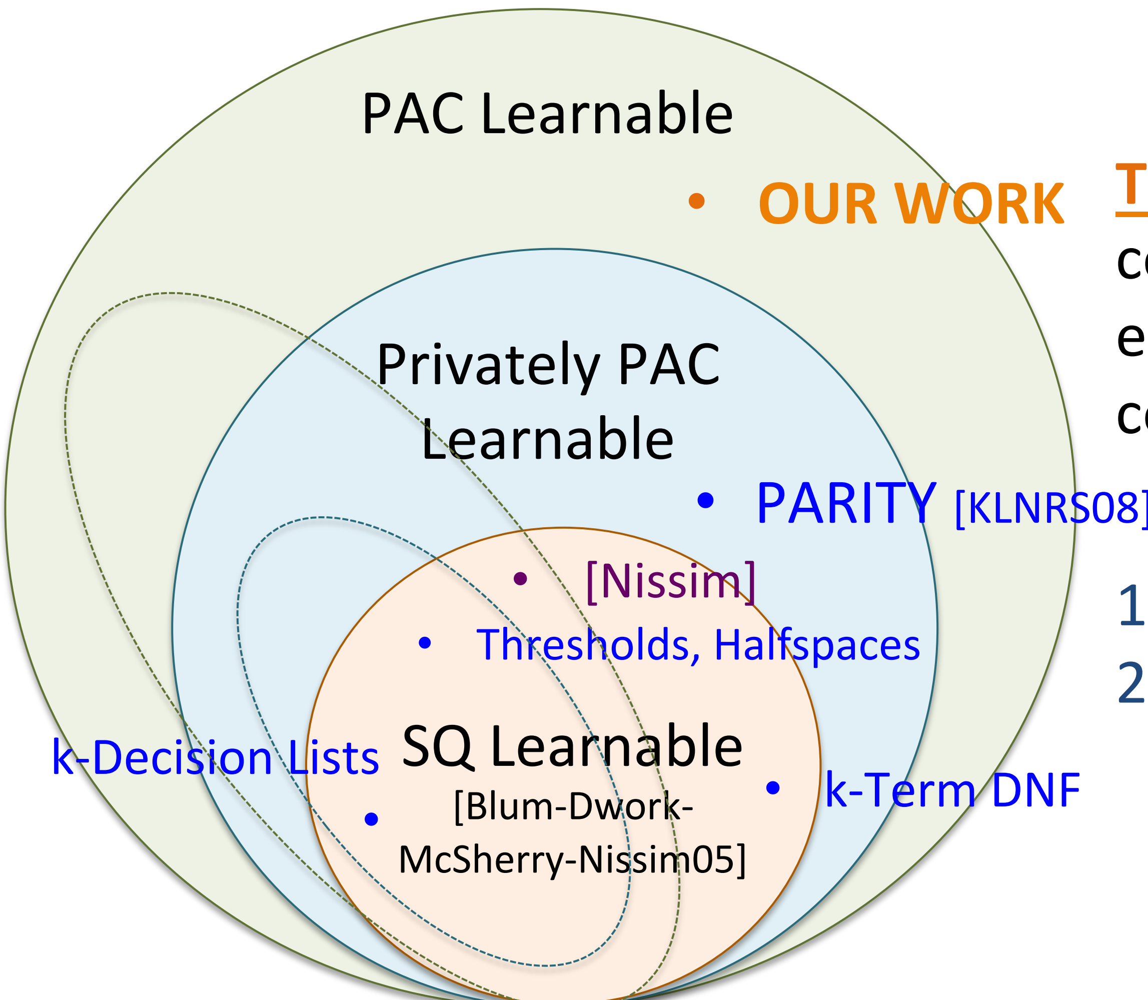


Evidence for a separation:

- Hardness of *representation-dependent* private learning [BKN10, Nissim]
- Private learning can require higher *sample complexity* [BKN10, BNS13, FX14, BNSV15, BNS16]
- Long tradition of privacy & learning lower bounds via crypto

◆Privacy ◆Accuracy ◆Complexity

What can be Learned Privately *and* Efficiently?

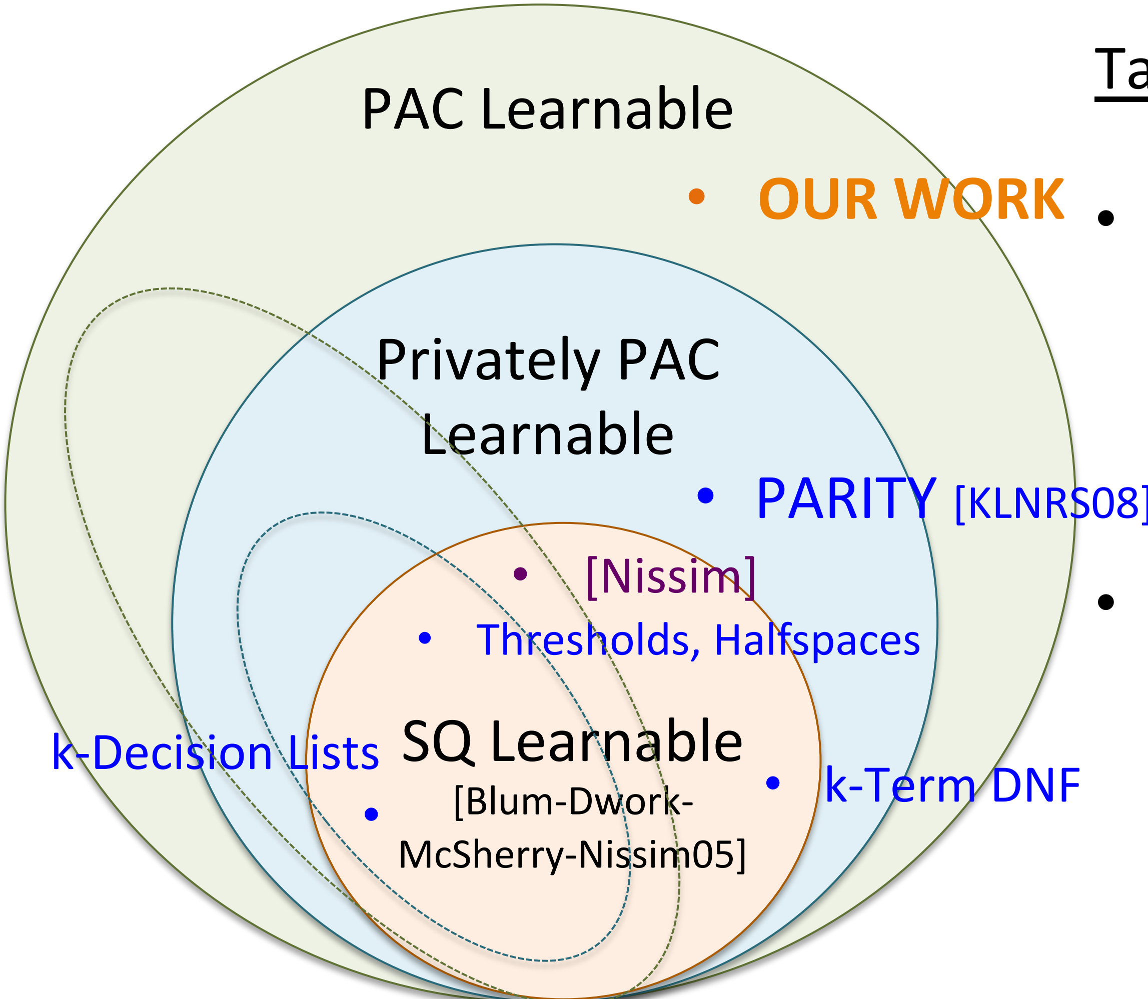


Thm: Assuming "strongly correct order-revealing encryption," there exists a concept class *C* that is

1) Efficiently PAC learnable
2) Hard to learn privately

◆Privacy ◆Accuracy ◆Complexity

What can be Learned Privately *and* Efficiently?

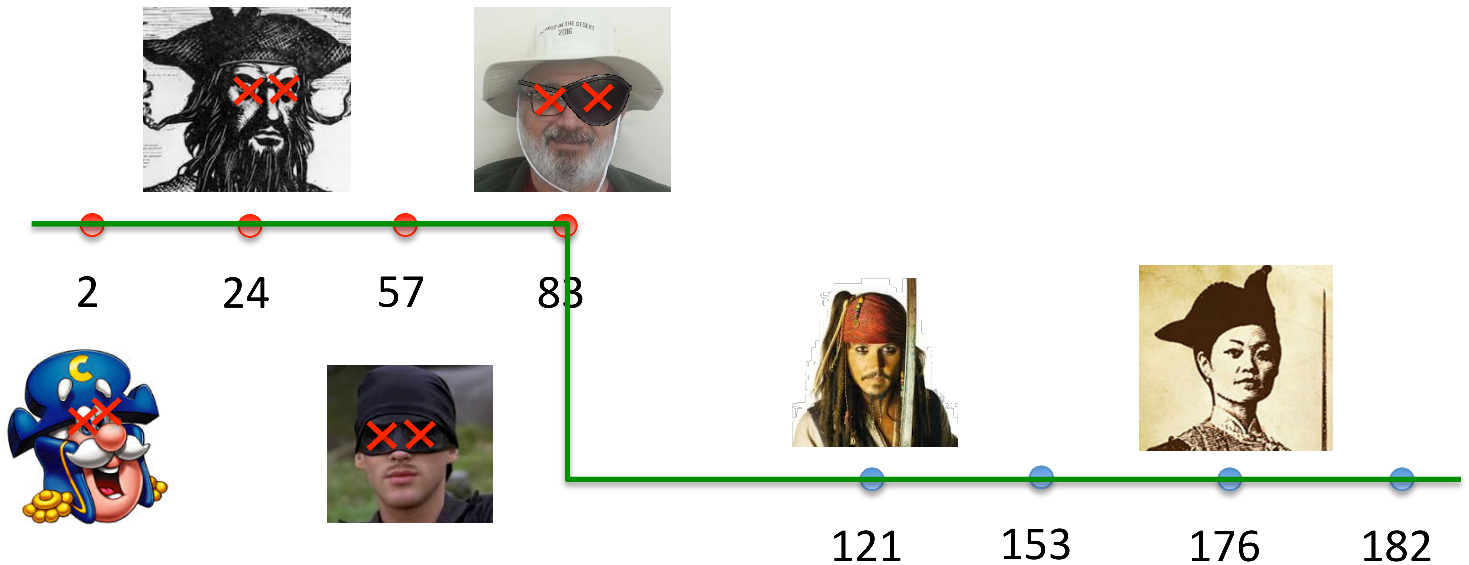


Takeaways

- "Learning from encrypted data" can still compromise differential privacy
- Separation between PAC and SQ learning w/o Gaussian elimination
  [cf. Feldman-Kanade12]

◆Privacy ◆Accuracy ◆Complexity

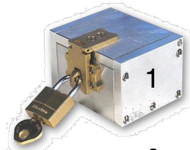
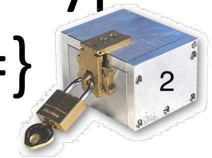
Our Separation



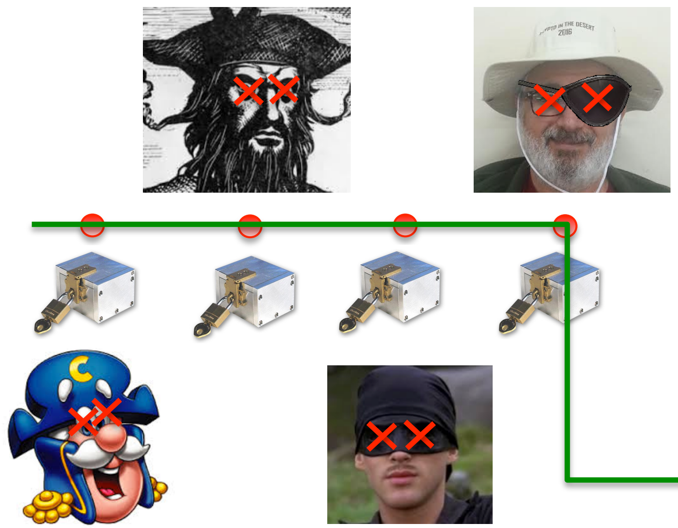
<u>Observation:</u>
Non-private learner only needs to compare the data

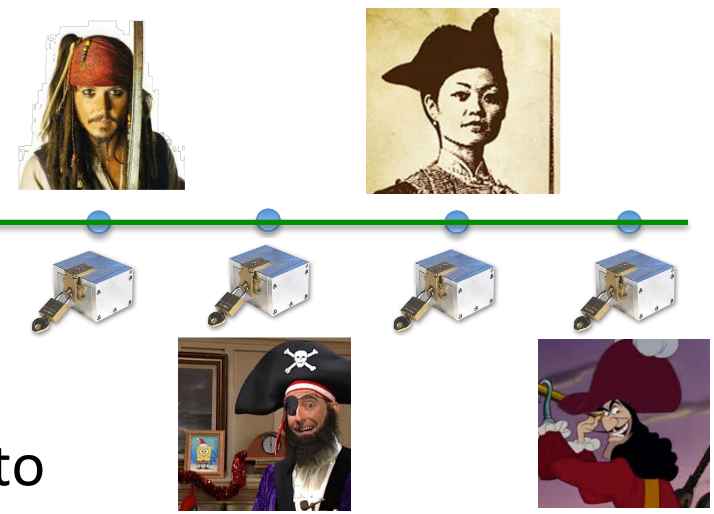
Our Separation

Order-Revealing Encryption:

Learn  {<, >, =} 

. . .but nothing else





<u>Observation:</u>

Non-private learner only needs to compare the data

Order-Revealing Encryption

[Boldyreva-Chenette-O'Neill11, Pandey-Rouselakis12]

x —Enc→ x (encrypted) ; x (encrypted) —Dec→ x

Enc,Dec secret

x (encrypted), y (encrypted) —Comp→ $x \le y$?

Comp public

IND-OCPA Security

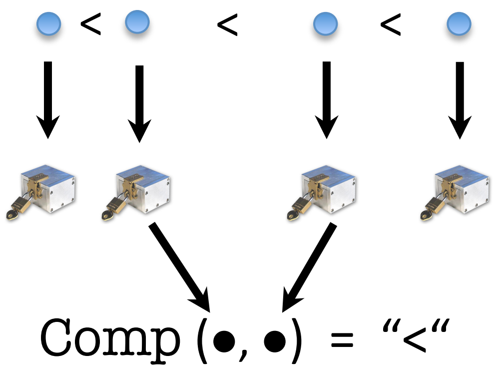
$x_1 < x_2 < \ldots < x_n$ → x_1, x_2, …, x_n (encrypted)

$\approx_c$

$y_1 < y_2 < \ldots < y_n$ → y_1, y_2, …, y_n (encrypted)

ORE vs. Order-Preserving Encryption

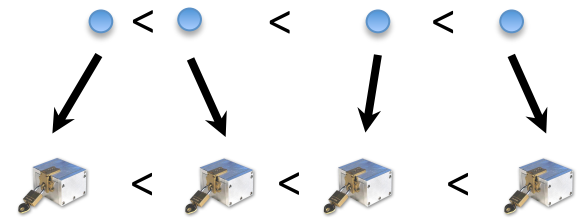
Order-Revealing



- Public Comp algorithm
- Known constructions ... strong assumptions
- **"Best possible" IND-OCPA security**

Order-Preserving

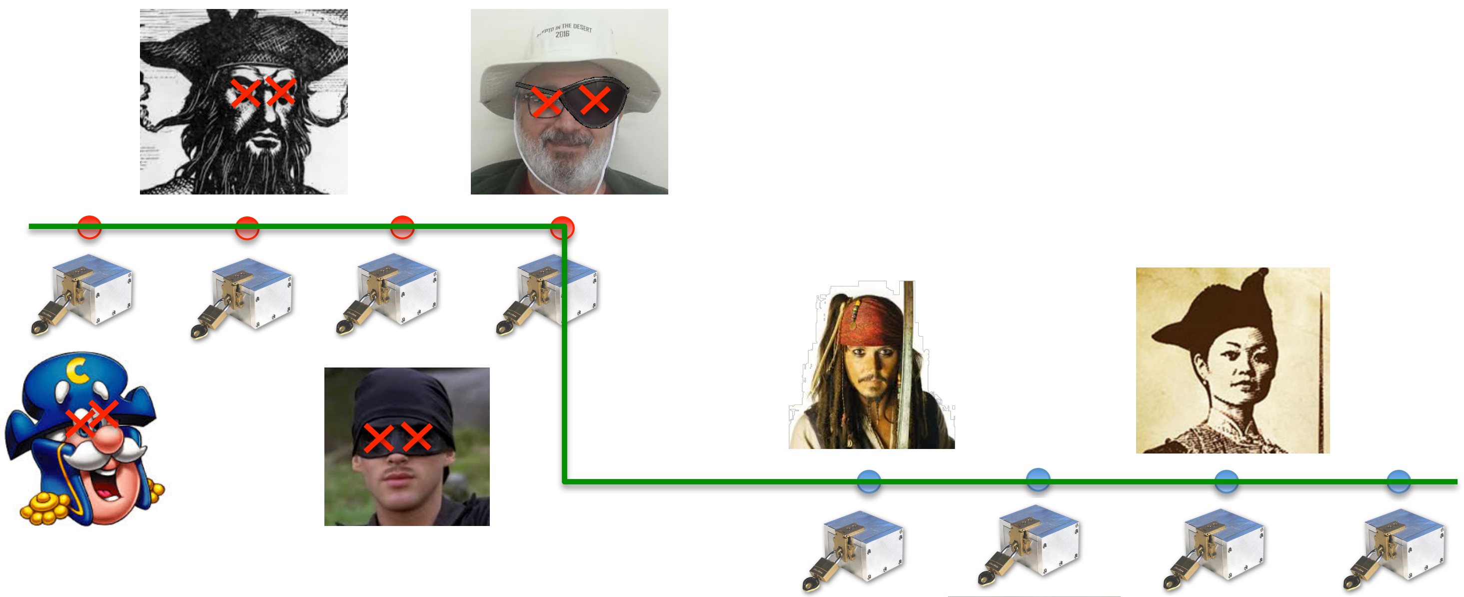
[Boldyreva-Chenette-Lee-O'Neill09]

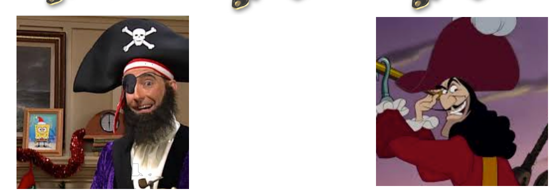


- Ciphertexts themselves ordered
- ...s under standard ...mptions
- Security unclear; necessarily leaks more than order

Crucial to our reduction

Our Separation





Domain X = ciphertext space

$\mathcal{C} = \{f_t(x) : \mathrm{Dec}(x) \leq t\ ?\}$

Things to prove:

1) $\mathcal{C}$ is PAC learnable   2) $\mathcal{C}$ is not privately learnable

Proof Ideas

1) PAC Learnability

~~Weak correctness~~

~~$\forall$ messages x, y: $\mathrm{Comp}(\mathrm{Enc}(x), \mathrm{Enc}(y)) = (x \leq y?)$~~

Strong correctness

$\forall$ ciphertexts c_0, c_1: $\mathrm{Comp}(c_0, c_1) = (\mathrm{Dec}(c_0) \leq \mathrm{Dec}(c_1)?)$

2) Hardness of Private Learning

Intuition: ORE forces learner to compare to a known example

Formally: Design an algorithm that "traces" an input example w.h.p.

(Conceptually analogous to [DNRRV09, Ullman13, BUV14, BZ15])

Is Our Assumption Reasonable?

- Constructions of weakly correct ORE:
  - iO [Garg-Gentry-Halevi-Raykova-Sahai-Waters13]
  - Functional encryption [GGHZ14+BS15, BLRSZZ15]

Multilinear Maps

- Can build strongly correct ORE from

  Weakly correct ORE + NIZKs [Groth-Ostrovsky-Sahai06]

Conclusions

- New source of hardness for private/SQ learning based on **order-revealing encryption**

Thank you!

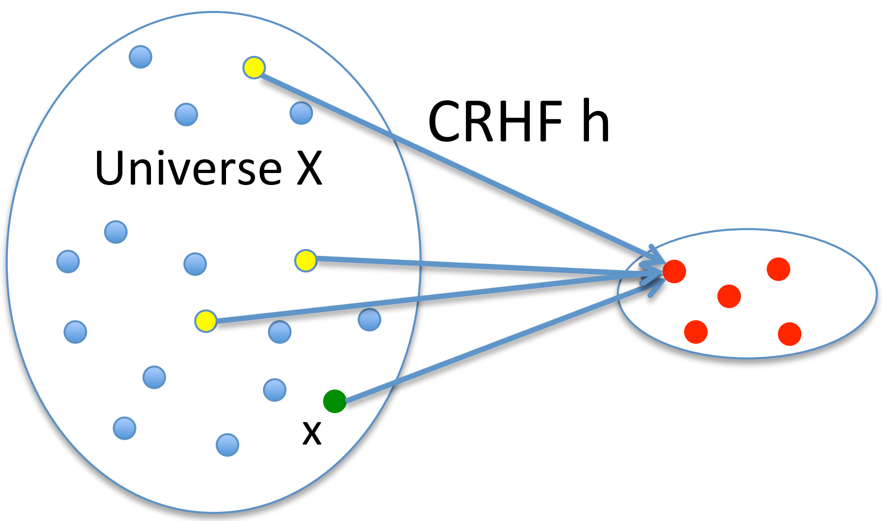
- Open questions:
  - Reduce to standard assumptions

  - Establish separation for "natural" learning problems

    [Ullman-Vadhan11, Daniely-Linial-ShalevShwartz14 et seq.]

Evidence for a Separation

$\mathcal{C}$ eff. PAC-learnable, but some *representation* of $\mathcal{C}$ is hard to learn privately [Nissim]



$\mathcal{C} = \mathcal{H} = \{f_x(y): h(x) = h(y)?\}$

Any positive example x is a representation of f_x
$\Rightarrow$ $\mathcal{C}$ is efficiently representation-learnable

Given positive examples, infeasible to find *new* rep.
$\Rightarrow$ Cannot privately learn a representation x