Challenges in Privacy-Preserving Data Analysis

Kamalika Chaudhuri

University of California, San Diego

Sensitive Data

Medical Records

Genetic Data

Search Logs

Google

AOL Violates Privacy

AOL Violates Privacy

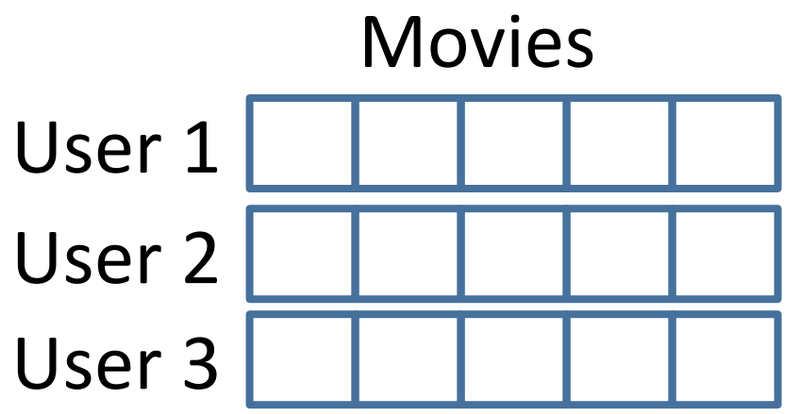
A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr.
Published: August 9, 2006

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.

No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 single men" to "dog that urinates on

Netflix Violates Privacy



2-8 movie-ratings and dates for Alice reveals:

Whether Alice is in the dataset or not

Alice's other movie ratings

High-dimensional Data is Unique

Example: UCSD Employee Salary Table

Position	Gender	Department	Ethnicity	Salary
Faculty	Female	CSE	SE Asian	-

One employee (Kamalika) fits description!

Simply anonymizing data is **unsafe!**

Disease Association Studies

Cancer

1.00										
.190	1.00									
.216	.251	1.00								
.186	.117	.047	1.00							
.154	.011	.170	.083	1.00						
.190	.140	.102	.095	.139	1.00					
.270	.215	.294	.248	.140	.141	1.00				
.101	.085	.170	.056	.234	.099	.175	1.00			
.239	.071	.163	.111	.161	.093	.199	.157	1.00		
.471	.117	.243	.094	.144	.123	.283	.216	.274	1.00	
.179	.202	.132	.094	.087	.159	.207	.108	.092	.294	1.00

Correlations

Healthy

1.00										
.141	1.00									
.099	.175	1.00								
.093	.199	.157	1.00							
.123	.283	.216	.274	1.00						
.159	.207	.108	.092	.294	1.00					
.088	.152	.075	.163	.156	.220	1.00				
.046	.161	.092	.072	.157	.143	.147	1.00			
.078	.392	.122	.229	.160	.172	.145	.177	1.00		
.045	.155	.135	.139	.110	.048	.126	.104	.169	1.00	
.178	.135	.102	.258	.314	.165	.147	.158	.131	.074	1.00

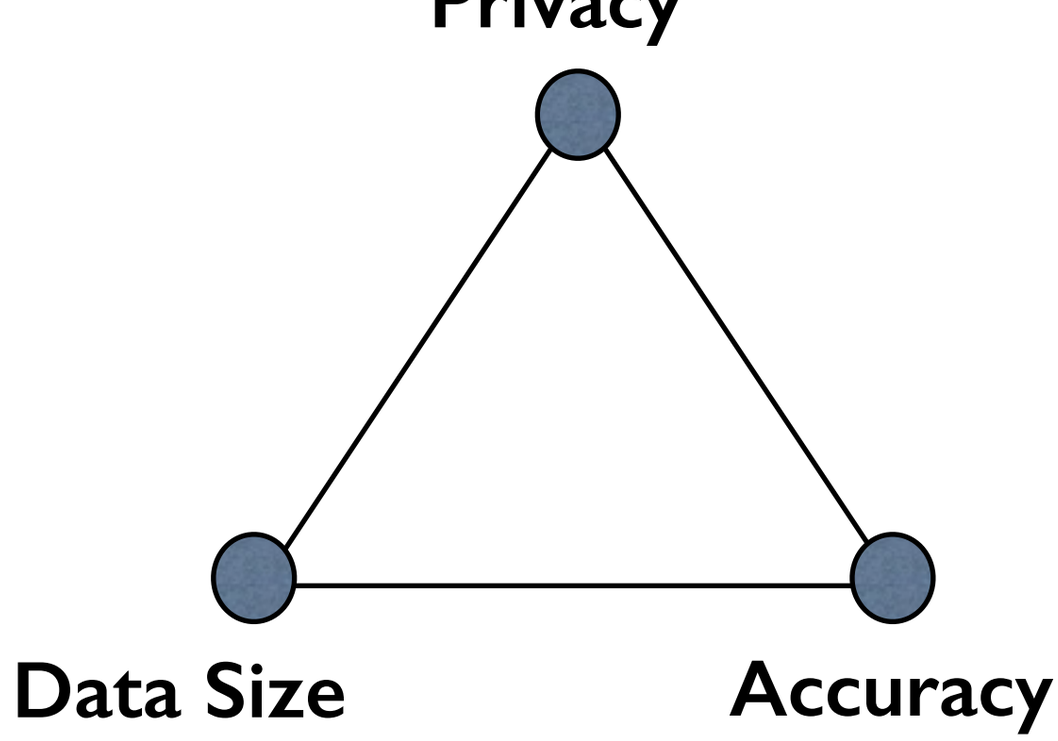
Correlations

Correlation (R^2 values), Alice's DNA reveals:

If Alice is in the **Cancer** set or **Healthy** set

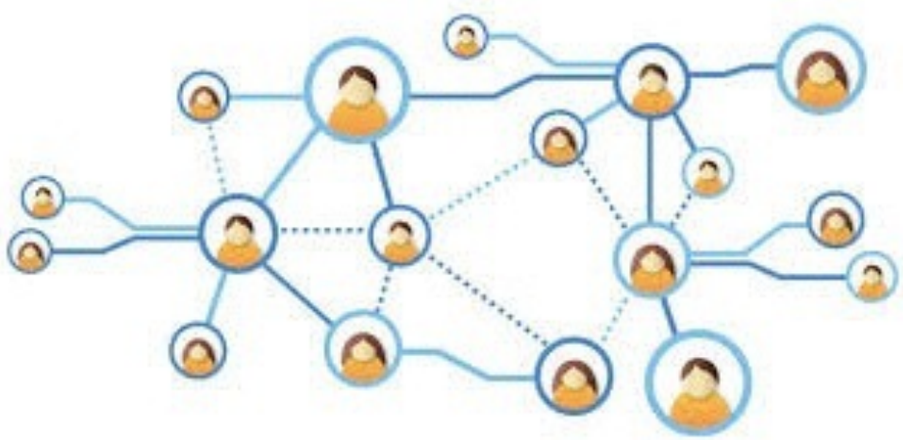
Simply anonymizing data is **unsafe!**

Statistics on small data sets is **unsafe!**

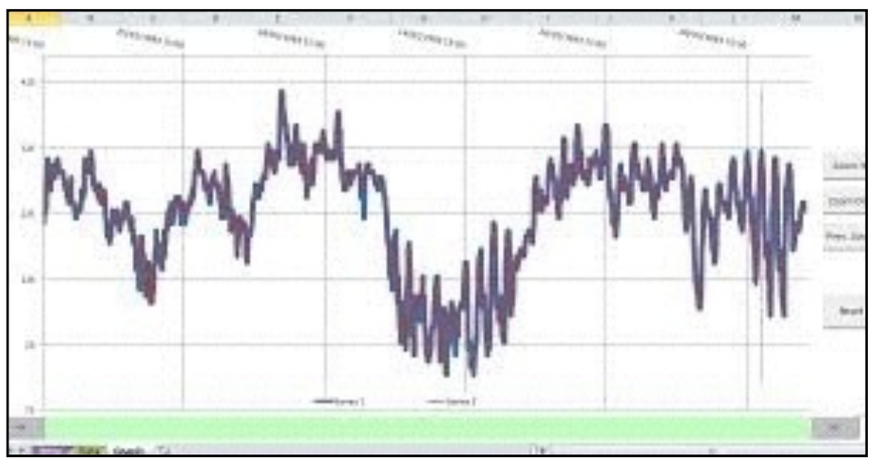


Correlated Data

User information
in social networks



Physical Activity
Monitoring



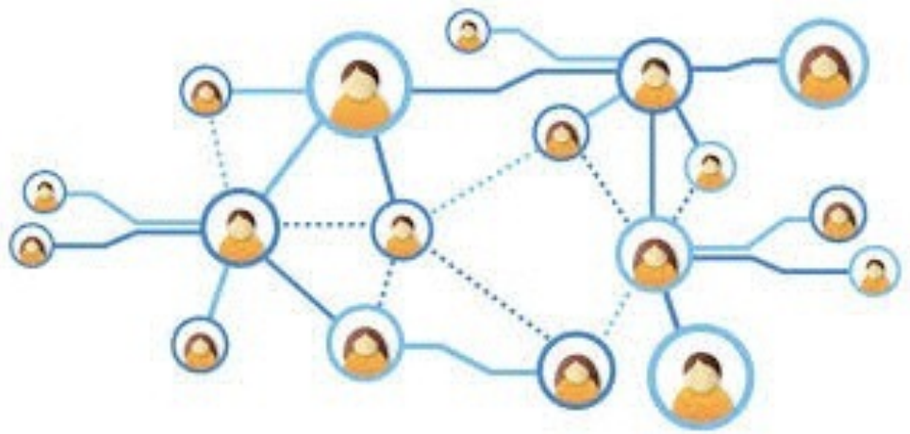
Why is Privacy Hard for Correlated data?

Correlation: neighbor's info leaks info on user

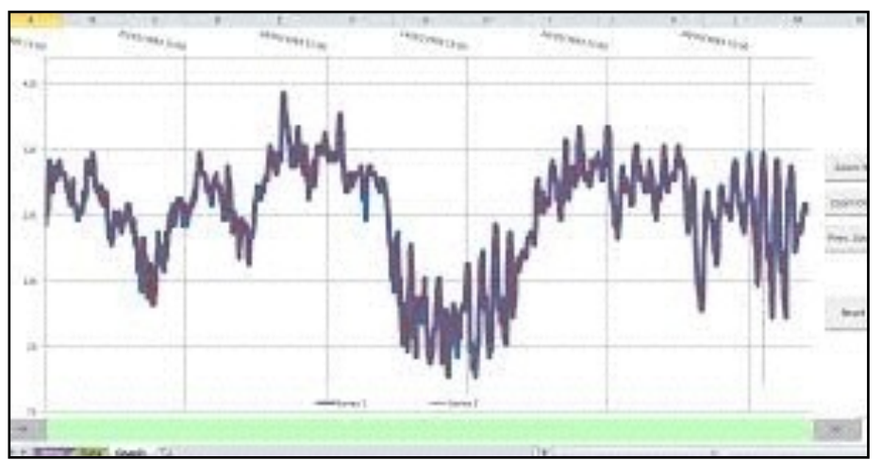
Why is Privacy Hard for Correlated data?

Correlation: neighbor's info leaks info on user

User information in social networks



Physical Activity Monitoring



Talk Agenda:

How do we **learn** from **sensitive** data while still preserving **privacy**?

How do we compute **statistics** on **correlated data** while still preserving **privacy**?

Talk Agenda:

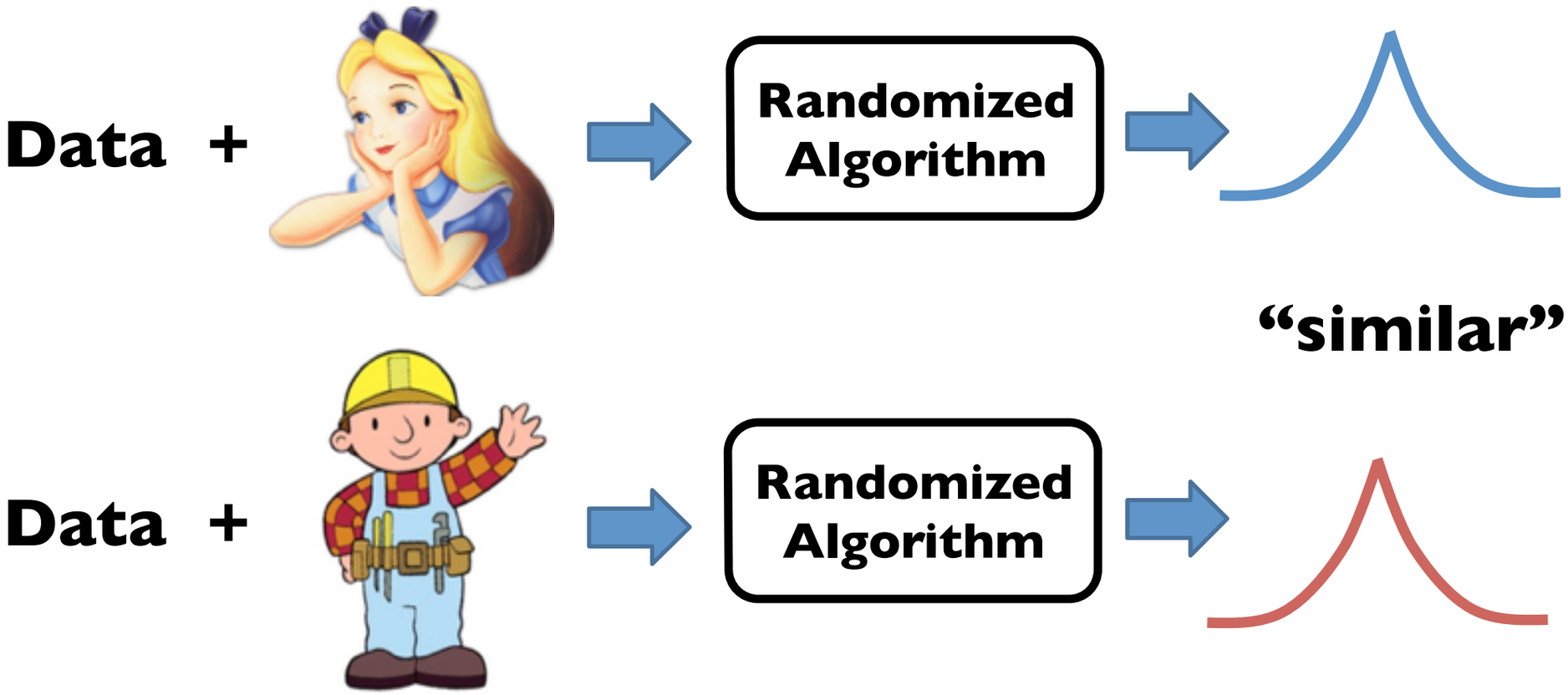
1. Privacy for Uncorrelated Data
    - How to define privacy
    - Privacy-preserving Classification

2. Privacy for Correlated Data
    - How to define privacy
    - Privacy-preserving Statistics

Talk Agenda:

1. Privacy for Uncorrelated Data
    - How to define privacy

Differential Privacy



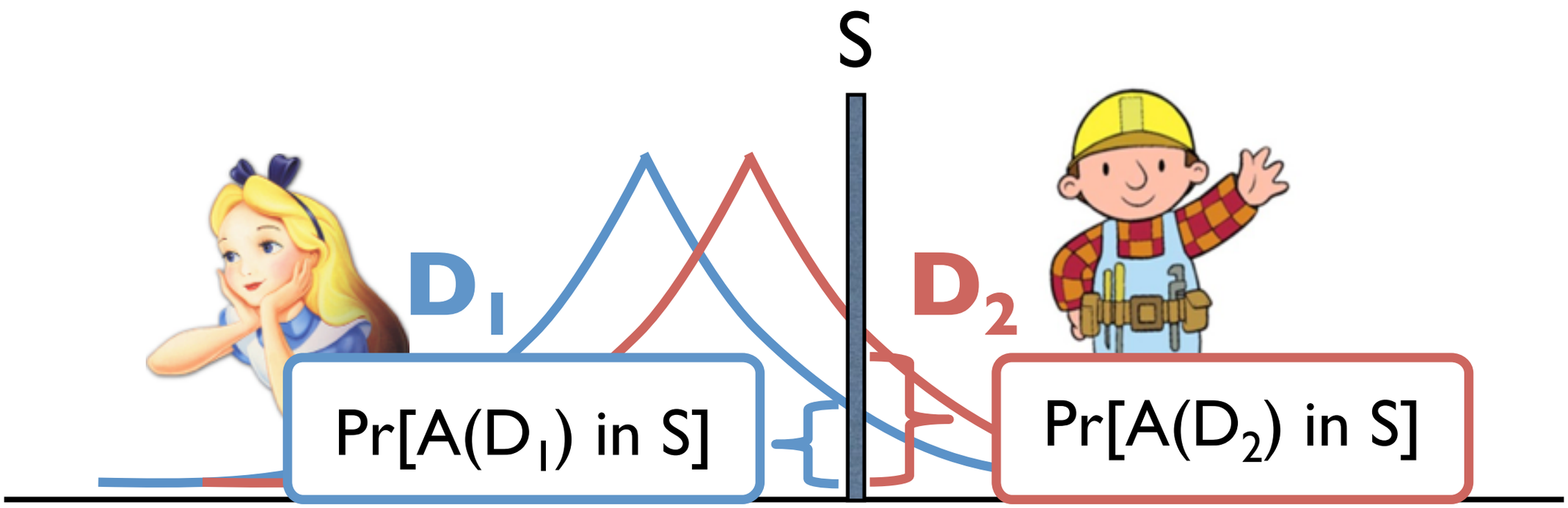
Participation of a single person does not change output

Differential Privacy: Attacker's View

Prior Knowledge + **Algorithm** Output on Data &  = **Conclusion** on 

Prior Knowledge + **Algorithm** Output on Data &  = **Conclusion** on 

Differential Privacy [DMNS06]



For all D_1, D_2 that differ in one person's value, any set S,
If A = α-private randomized algorithm, then:

$$\Pr(A(D_1) \in S) \leq e^{\alpha} \Pr(A(D_2) \in S)$$

Differential Privacy

1. **Provably strong notion of privacy**

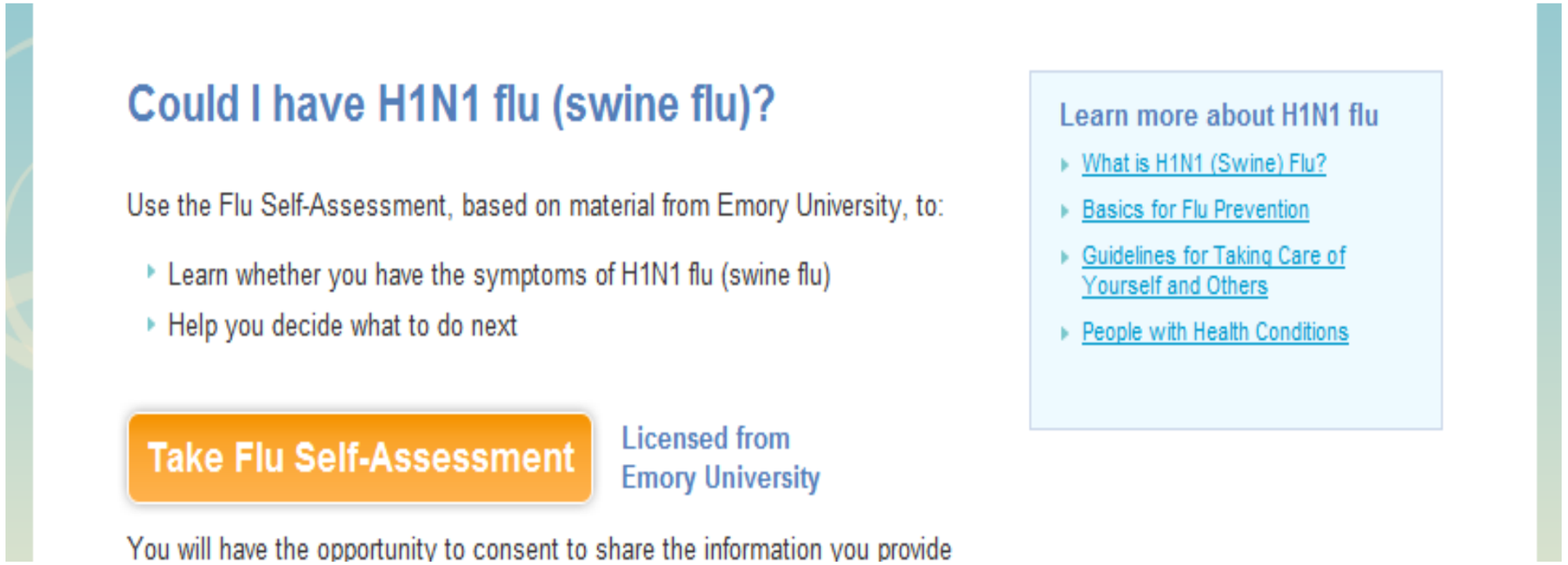
2. **Good approximations for many functions**

   e.g, means, histograms, etc.

Talk Agenda:

1. Privacy for Uncorrelated Data
    - How to define privacy
    - Privacy-preserving Classification

A Classification Problem: Flu Test

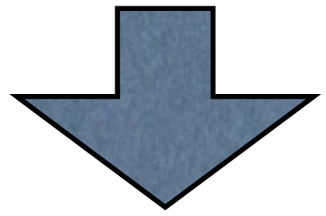


Predicts flu or not, based on patient symptoms

Trained on sensitive patient data

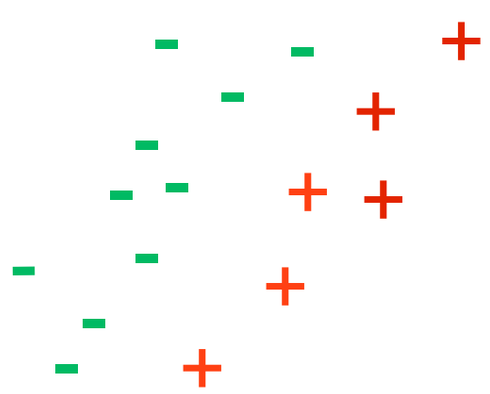
From Attributes to Labeled Data

Yes	No	99F	No
Sore Throat	Fever	Temperature	Flu?



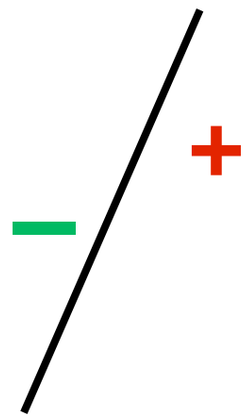
1	0	99	–
Data			Label

Classifying Sensitive Data



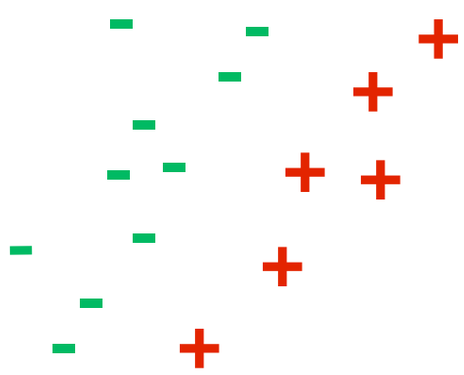
Private Data

Classification Algorithm

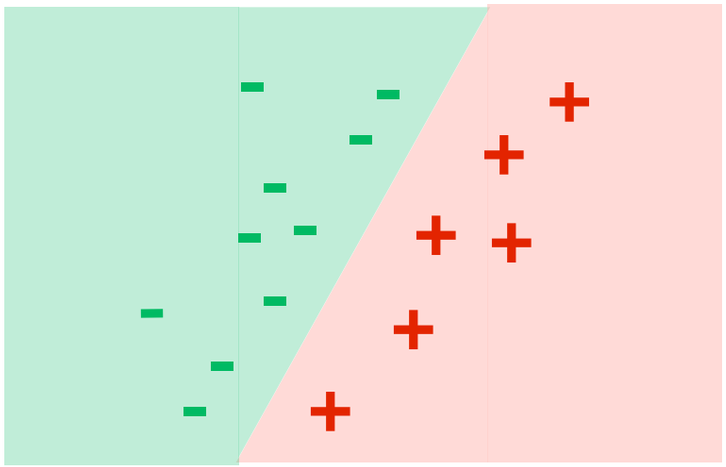


Public Classifier

Linear Classification

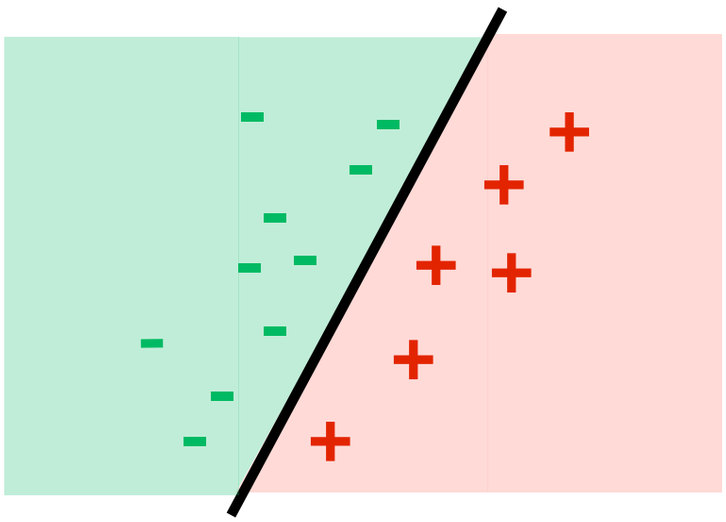


Linear Classification



Distribution P over
labelled examples

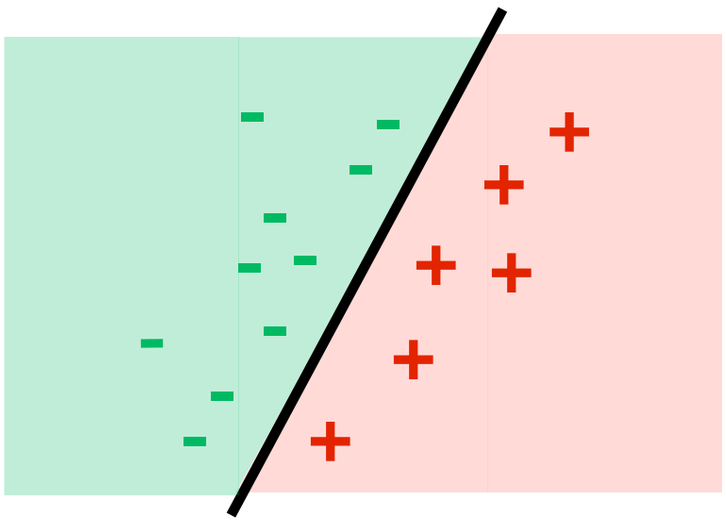
Linear Classification



Distribution P over labelled examples

Goal: Find a vector w that separates + from - for most points from P

Linear Classification



Distribution P over labelled examples

Goal: Find a vector w that separates + from - for most points from P

Key: Find a simple model to fit the samples

Empirical Risk Minimization

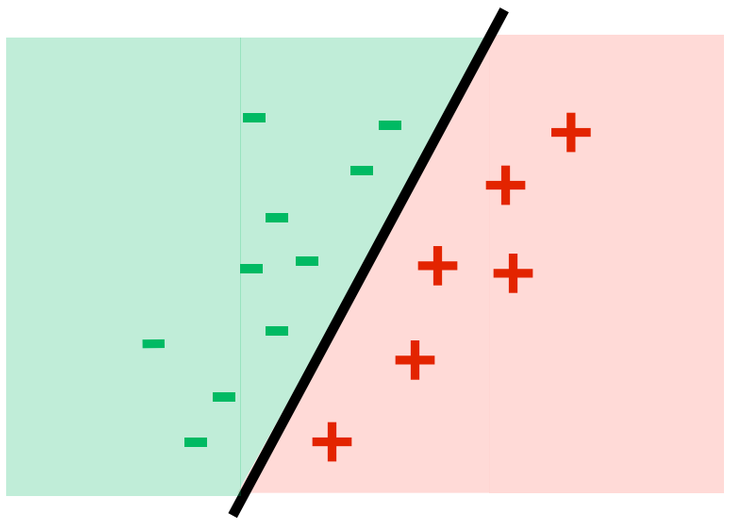
Goal: Labeled data (x_i, y_i), find w minimizing:

$$\frac{1}{2}\lambda\|w\|^2 \quad + \quad \frac{1}{n}\sum_{i=1}^{n} L(y_i w^T x_i)$$

Regularizer
(Model Complexity)

Risk
(Training Error)

Some Examples



Risk: Hinge loss
Optimizer: Support vector machines (SVM)

Risk: Logistic loss
Optimizer: Logistic regression

ERM with Privacy

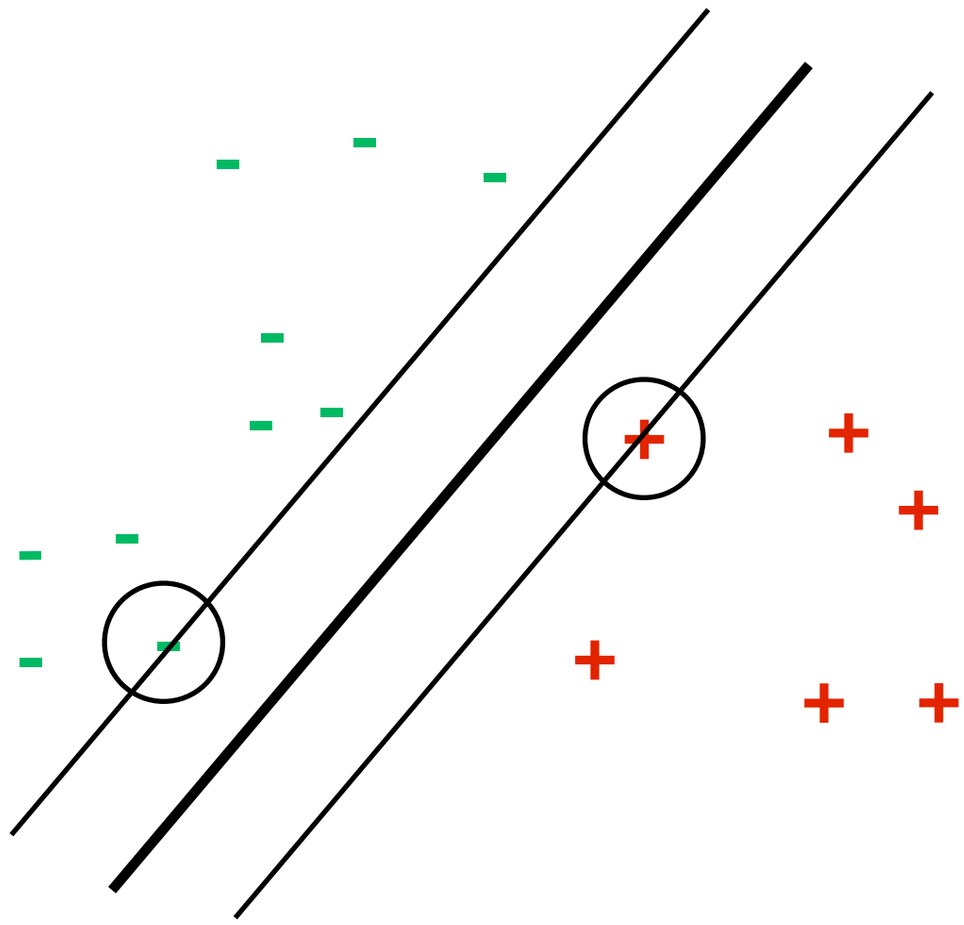
Given: labeled data (x_i, y_i),

Find: vector w that is:

(Private)
Is private w.r.t training data

(Accurate)
Approximately minimizes the regularized risk

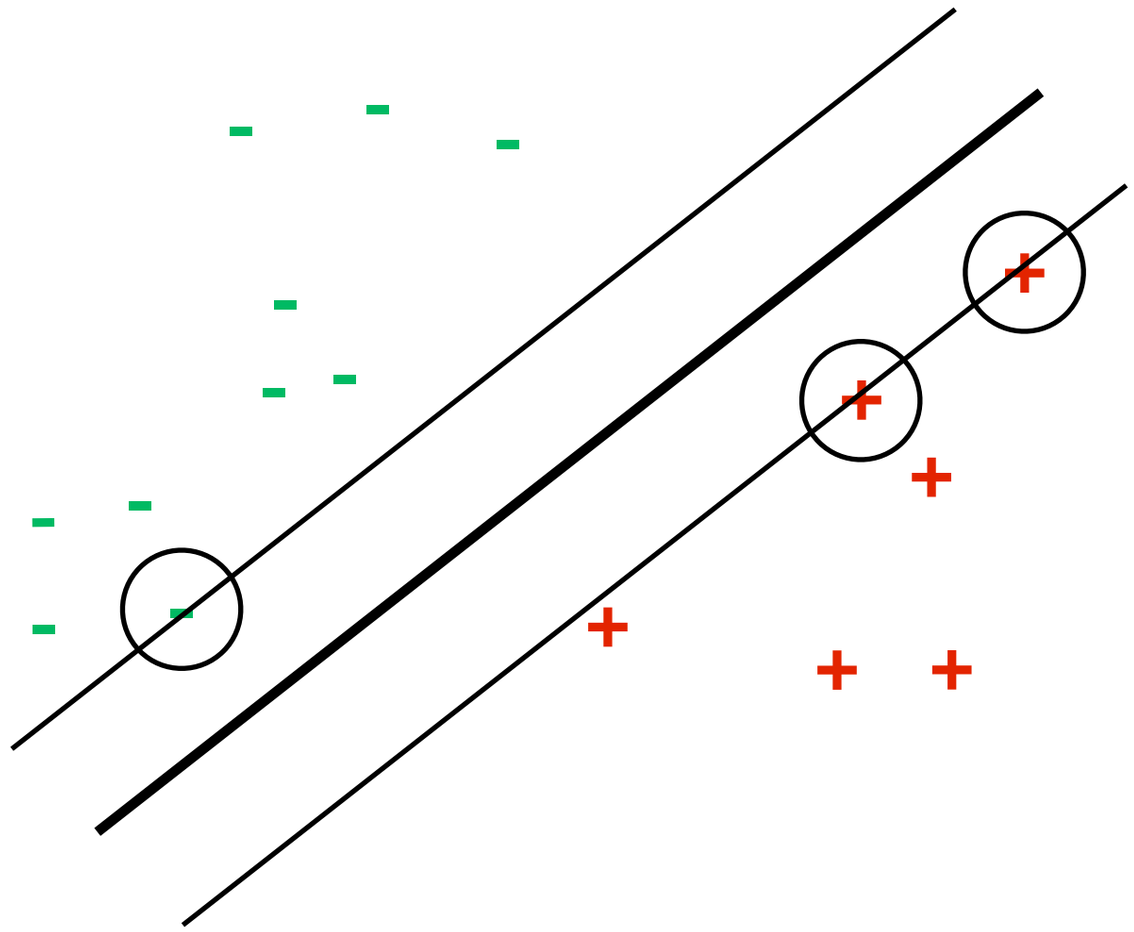
Why is ERM not private for SVM?



SVM solution is a combination of support vectors

If one support vector moves, solution changes

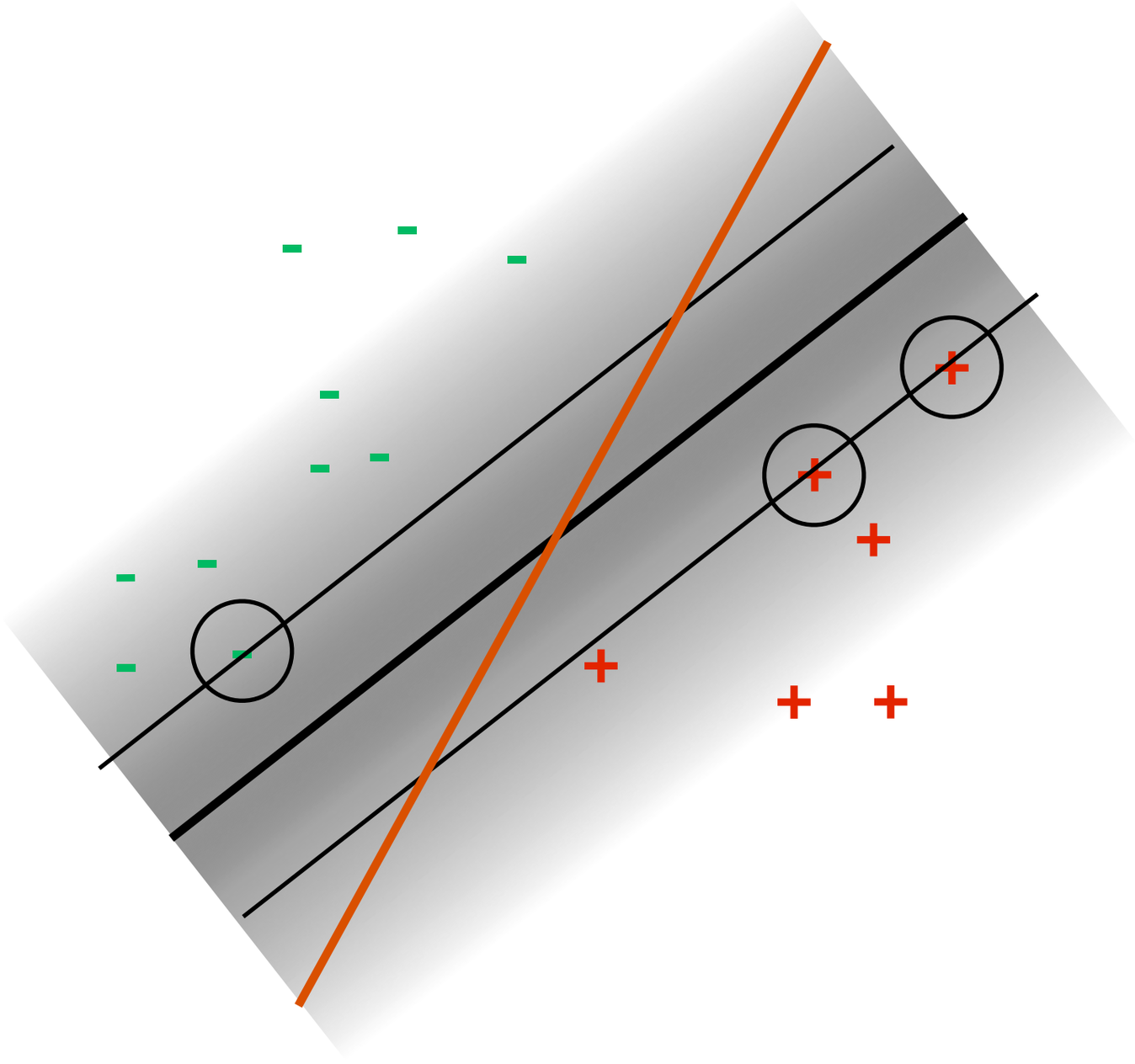
Why is ERM not private for SVM?



SVM solution is a combination of support vectors

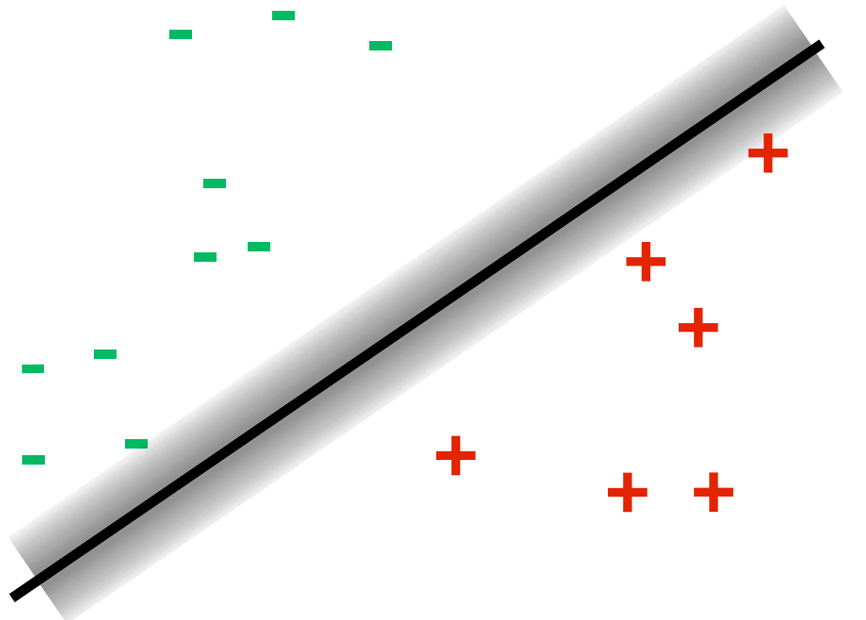
If one support vector moves, solution changes

How to make ERM private?

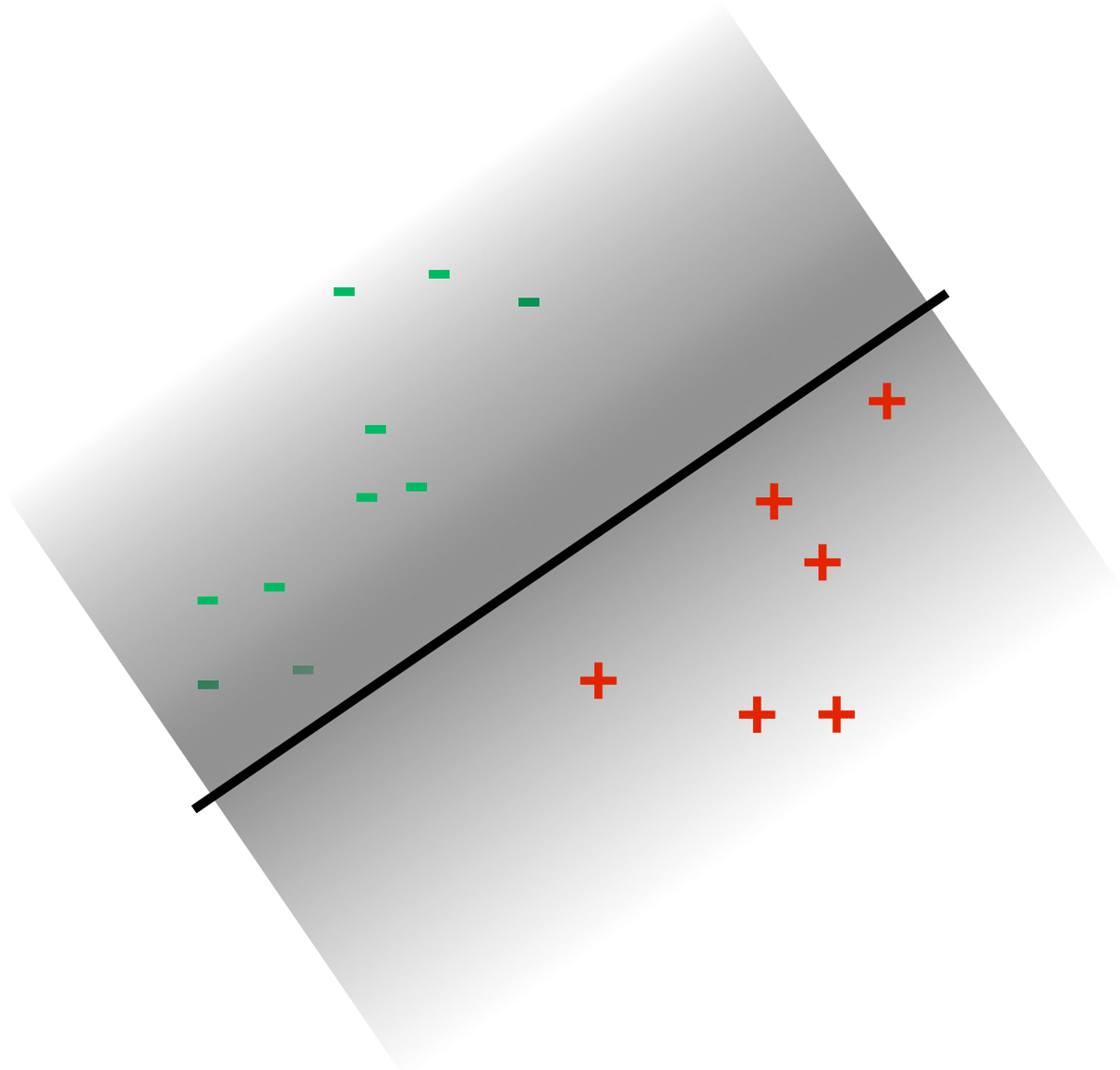


Pick w from distribution near the optimal solution

Privacy vs. Accuracy



Too little privacy



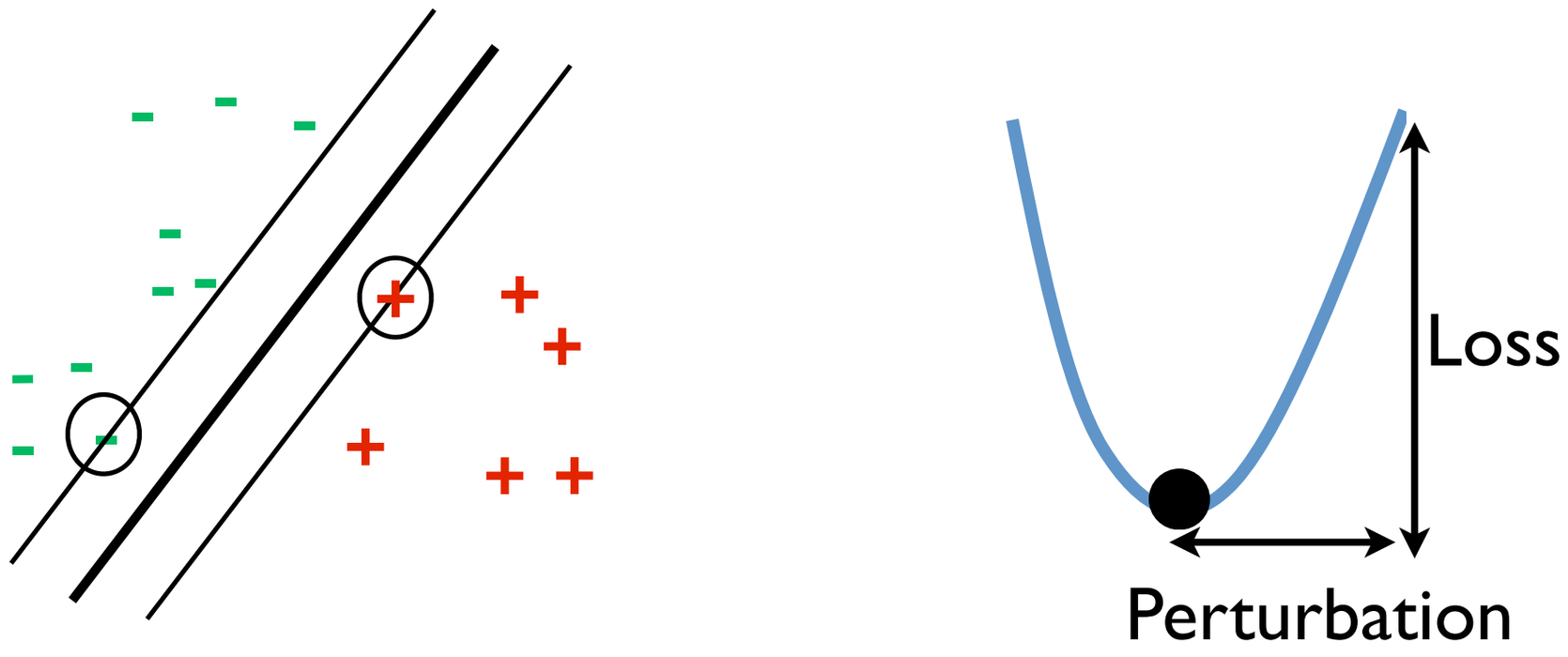
Too little accuracy

Pick distribution that gives **privacy** and **accuracy**

Privacy-preserving Classification

1. ERM with privacy
2. Algorithm

Properties of Real Data



Optimization surface is very steep in some directions

High loss if perturbed in those directions

Insight: Perturb optimization surface and then **optimize**

Empirical Risk Minimization

Goal: Labeled data (x_i, y_i), find w minimizing:

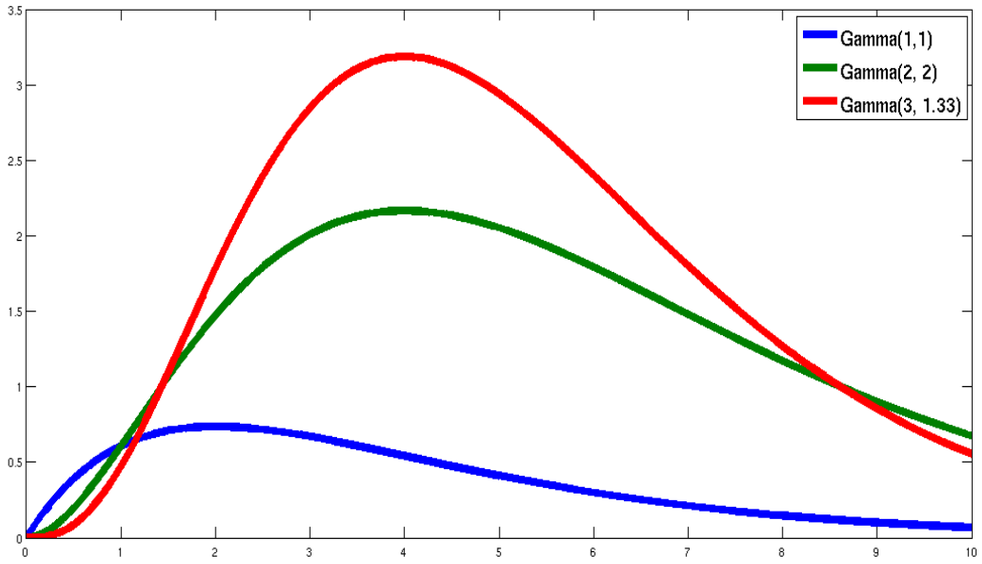
$$\frac{1}{2}\lambda\|w\|^2 \quad + \quad \frac{1}{n}\sum_{i=1}^{n} L(y_i w^T x_i) \quad + \quad \frac{1}{n} b^\top w$$

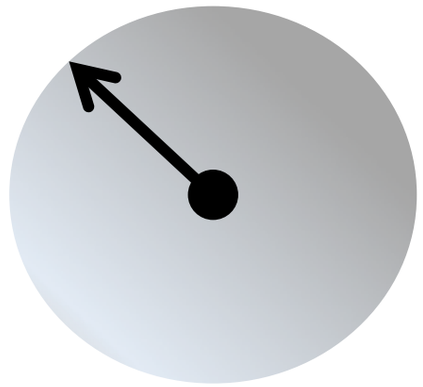
Regularizer
(Model Complexity)

Risk
(Training Error)

Perturbation
(Privacy)

Algorithm: Perturbation





Perturbation b drawn from:

Magnitude: Drawn from $\Gamma(d, 1/\alpha)$

Direction: Uniformly at random

Privacy-preserving Classification

1. ERM with privacy
2. Algorithm
3. Analytical guarantees

Privacy Guarantees

Algorithm: Given labeled data (x_i, y_i), find w to minimize:

$$\frac{1}{2}\lambda\|w\|^2 \quad + \quad \frac{1}{n}\sum_{i=1}^{n} L(y_i w^T x_i) \quad + \quad \frac{1}{n} b^\top w$$

Theorem: If L is convex and doubly-differentiable with $|L'(z)| \leq 1$ and $|L''(z)| \leq c$ then Algorithm is $\alpha + 2\log\left(1 + \frac{c}{n\lambda}\right)$-differentially private

Privacy Guarantees

Algorithm: Given labeled data (x_i, y_i), find w to minimize:

$$\frac{1}{2}\lambda\|w\|^2 \quad + \quad \frac{1}{n}\sum_{i=1}^{n} L(y_i w^T x_i) \quad + \quad \frac{1}{n} b^\top w$$

L = Logistic Loss ⇒ **Private Logistic Regression**

L = Huber Loss ⇒ **Private SVM**

(Hinge Loss is not differentiable)

Measure of Accuracy

Number of samples for error ϵ

(Fewer samples implies higher accuracy)

Sample Requirement

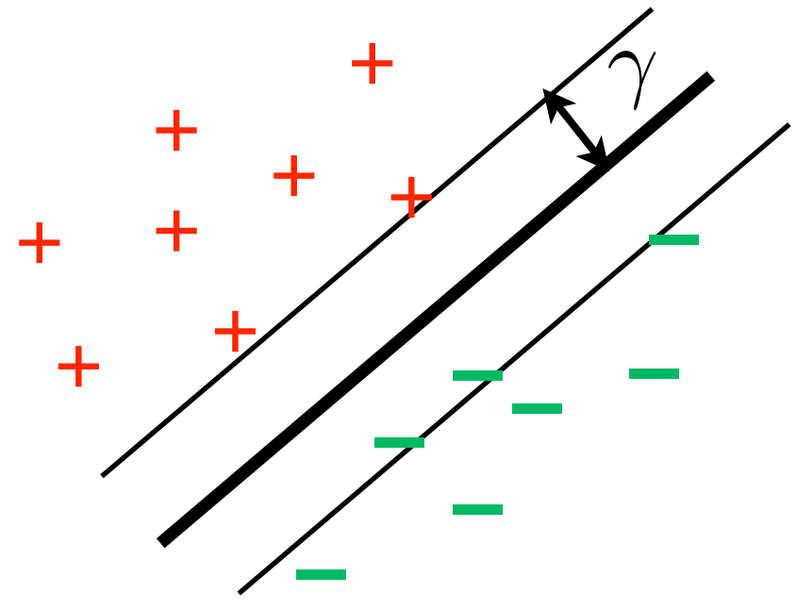
d : #dimensions

γ : margin

α : privacy

ϵ : error

$\gamma, \alpha, \epsilon < 1$



Normal SVM: $1/\epsilon^2\gamma^2$

Our Algorithm: $1/\epsilon^2\gamma^2 + d/\epsilon\alpha\gamma$

Standard Method: $1/\epsilon^2\gamma^2 + d/\epsilon^{3/2}\alpha\gamma$

Privacy-preserving Classification

1. ERM with privacy
2. Algorithm
3. Analytical guarantees
4. Evaluation

Experiments

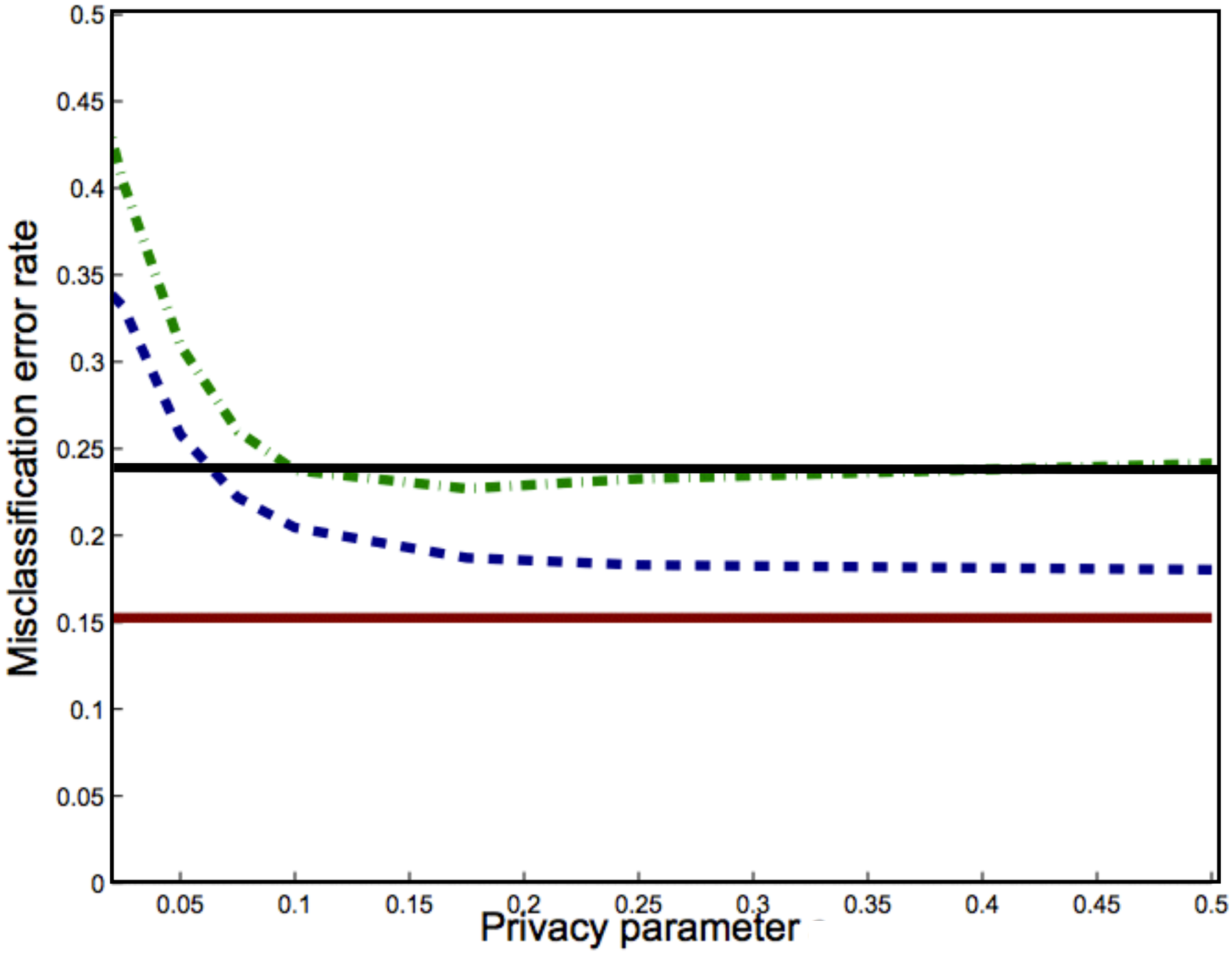
UCI Adult: Census/Income Data

Demographic dataset of size 47K

105 dimensions after preprocessing

Task: Predict if income above/below 50K

Results: SVM



Non-private

Standard Method

Our Method

Predict majority label

Experiments

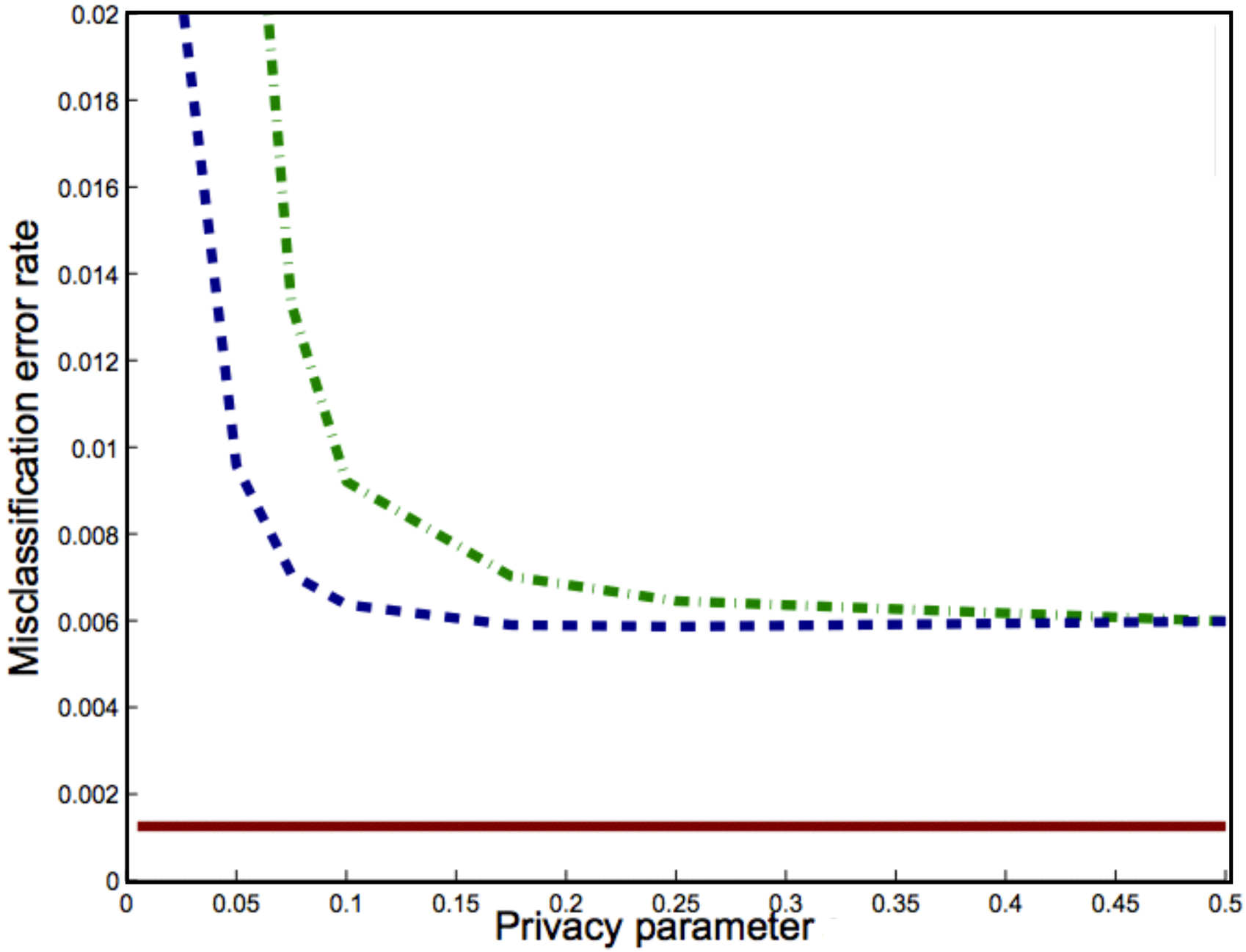
KDDCup99: Intrusion detection data

70K network connections

116 dimensions after preprocessing

Task: Predict if connection is malicious or not

Results: SVM



Non-private

Standard Method

Our Method

Privacy-preserving Classification

1. ERM with privacy
2. Algorithm
3. Analytical guarantees
4. Evaluation

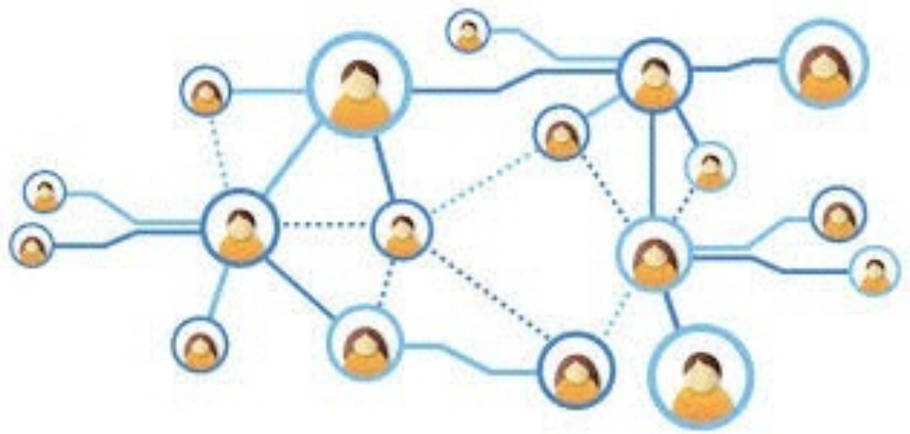
Talk Agenda:

1. Privacy for Uncorrelated Data
    - How to define privacy
    - Privacy-preserving Classification
2. Privacy for Correlated Data

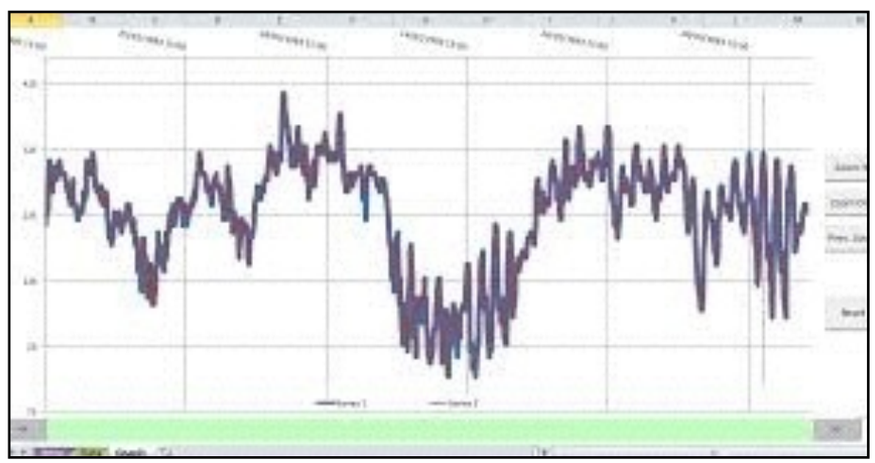
Why is Privacy Hard for Correlated data?

Correlation: neighbor's info leaks info on user

User information in social networks



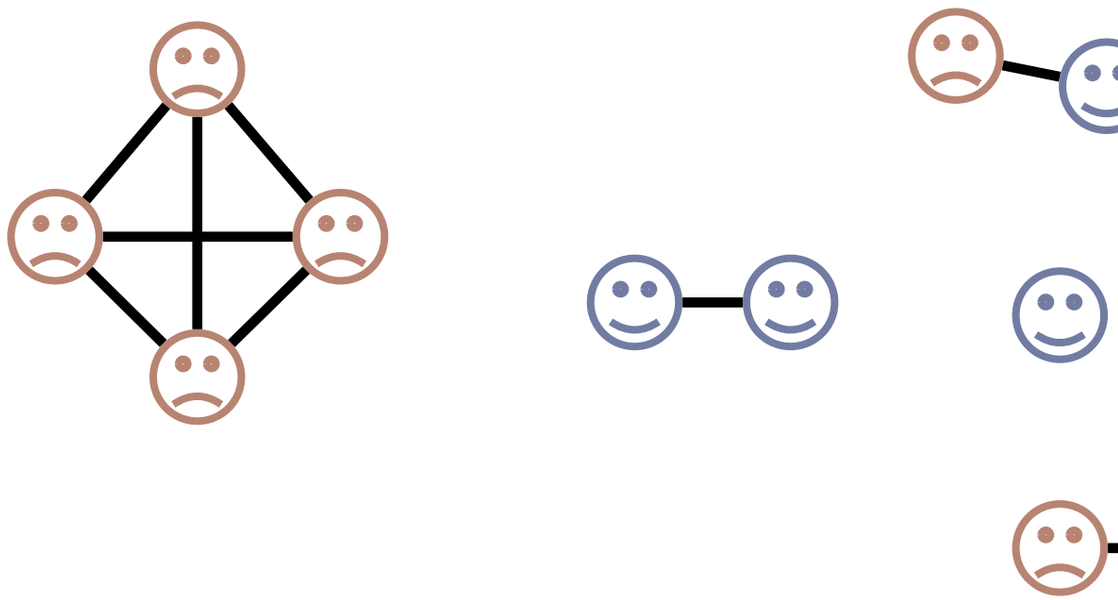
Physical Activity Monitoring



Why is Differential Privacy not Enough for Correlated data?

DP for correlated data

Example: $D = (x_1, .., x_n)$, $x_i = 1$ if i has flu

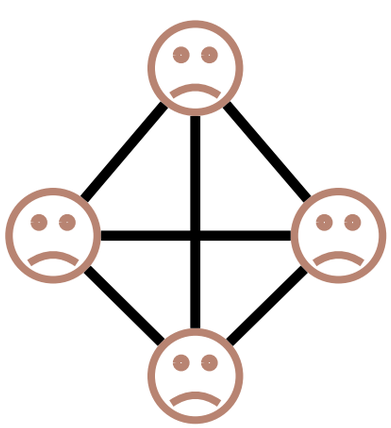


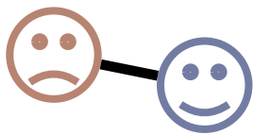
Interaction Network

Goal: (1) Publish #people with flu in D
(2) Prevent adversary from knowing who has flu

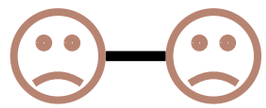
DP for correlated data

Example: D = $(x_1, .., x_n)$, x_i = 1 if i has flu





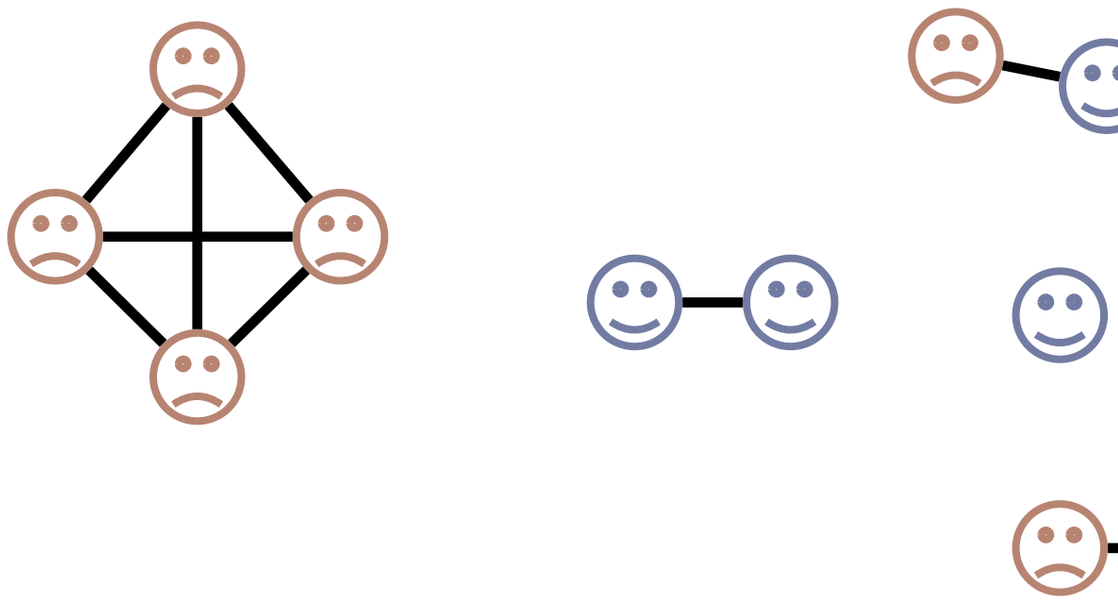
Interaction Network



1-DP: Output #people with flu + noise with stdev 1

DP for correlated data

Example: $D = (x_1, .., x_n)$, $x_i = 1$ if i has flu



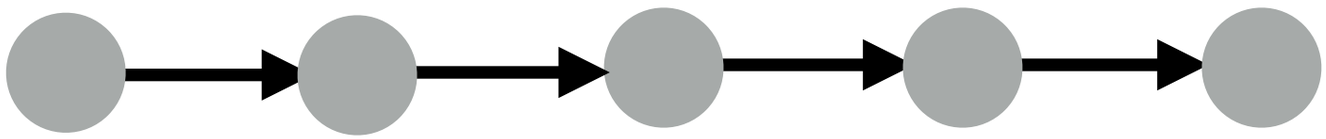
Interaction Network

1-DP: Output #people with flu + noise with stdev 1

Not enough for privacy of people in connected components!

DP for correlated data

Example: $D = (x_1, .., x_t)$, x_i = activity at time t

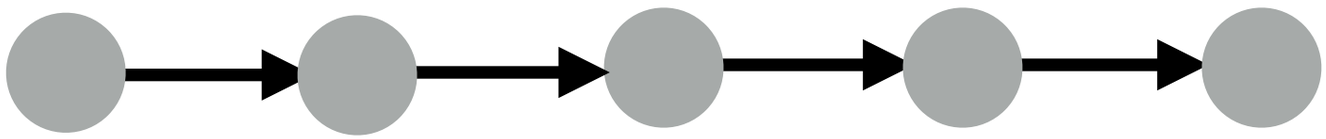


Correlation Network

Goal: (1) Publish activity histogram
(2) Prevent adversary from knowing activity at t

DP for correlated data

Example: $D = (x_1, .., x_t)$, x_i = activity at time t

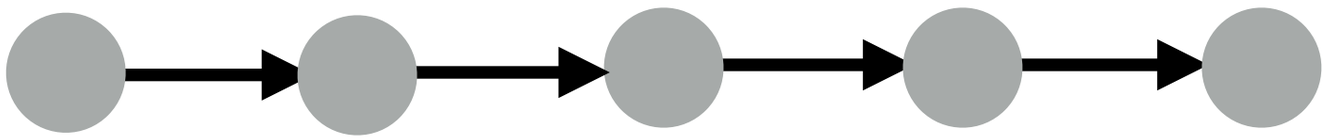


Correlation Network

1-DP: Output histogram of activities + noise with stdev 1

DP for correlated data

Example: D = (x_1, .., x_t), x_i = activity at time t



Correlation Network

1-DP: Output histogram of activities + noise with stdev 1

Not enough - activities across time are highly correlated!

Talk Agenda:

1. Privacy for Uncorrelated Data
    - How to define privacy
    - Privacy-preserving Classification

2. Privacy for Correlated Data
    - How to define privacy

Pufferfish Privacy: Components

Secrets S: Information to be protected

e.g: Alice's age is 25, Bob has a disease

Secret Pairs Q: Pairs of secrets to be indistinguishable

e.g: (Alice's age is 25, Alice's age is 40),
(Bob is in dataset, Bob is not in dataset)

Distribution Class Θ: Set of distributions that can plausibly generate the data

e.g: disease is passed on w.p. [0.1, 0.9]

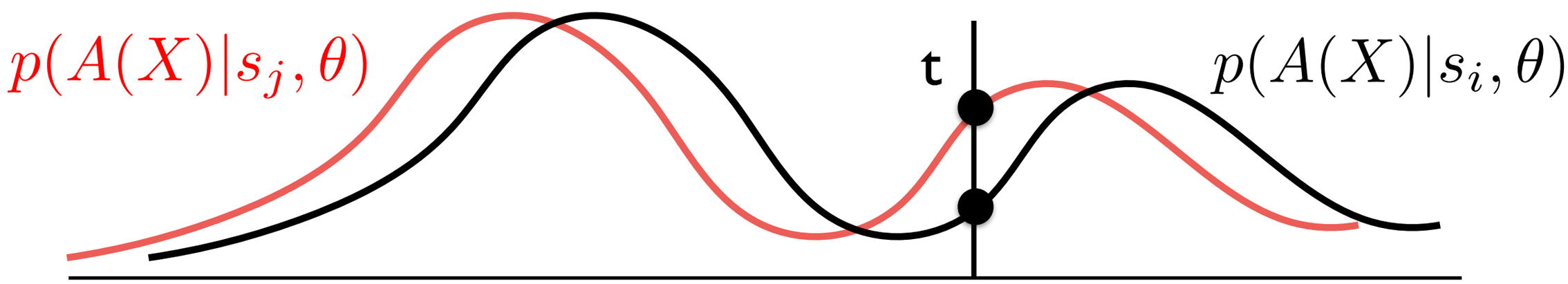
Distribution Class models correlation!

Pufferfish Privacy

An algorithm A is α-Pufferfish private with parameters (S, Q, Θ) if for all (s_i, s_j) in Q, for all $\theta \in \Theta$, $X \sim \theta$, all t,

$$\frac{p_{\theta,A}(A(X) = t | s_i, \theta)}{p_{\theta,A}(A(X) = t | s_j, \theta)} \leq e^{\alpha}$$

whenever $P(s_i|\theta), P(s_j|\theta) > 0$

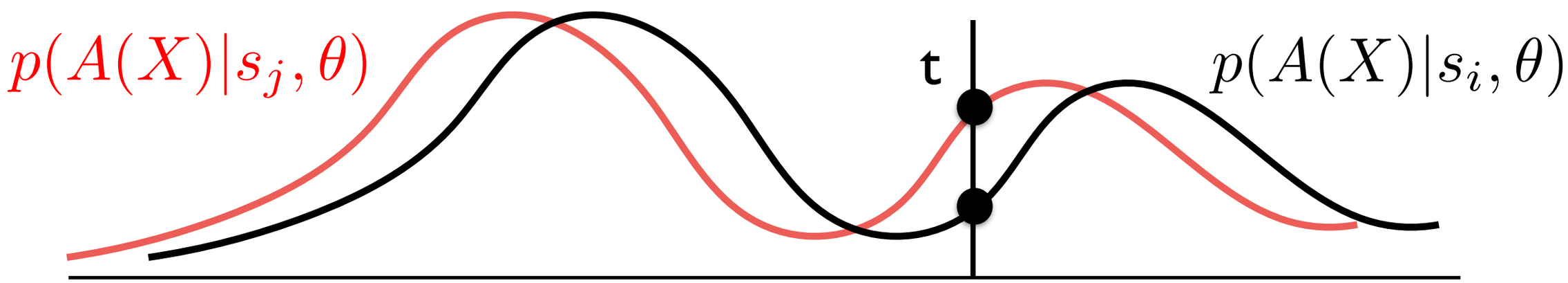


Pufferfish Privacy

An algorithm A is α-Pufferfish private with parameters (S, Q, Θ) if for all (s_i, s_j) in Q, for all $\theta \in \Theta$, $X \sim \theta$, all t,

$$e^{-\alpha} \leq \frac{p_{\theta,A}(s_i|A(X) = t, \theta)}{p_{\theta,A}(s_j|A(X) = t, \theta)} \Big/ \frac{p_\theta(s_i|\theta)}{p_\theta(s_j|\theta)} \leq e^{\alpha}$$

whenever $P(s_i|\theta), P(s_j|\theta) > 0$



Knowing A does not affect adversary's belief on s_i vs. s_j

Pufferfish generalizes DP

Theorem: α-DP is equivalent to α-Pufferfish privacy with parameters (S, Q, Θ), where

Q:= { (i in data with value a, i in data with value b), $\forall i, a, b$ }

U { (i in data with value a, j in data), $\forall i \neq j, a$ }

Θ := Each individual is **independent**

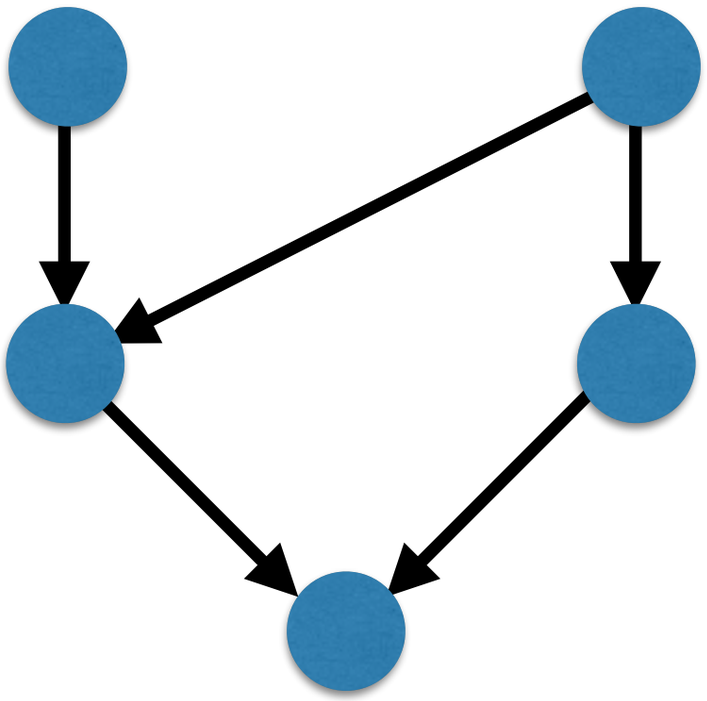
There can be no utility if Θ allows any arbitrary correlation!

Talk Agenda:

1. Privacy for Uncorrelated Data
    - How to define privacy
    - Privacy-preserving Classification

2. Privacy for Correlated Data
    - How to define privacy
    - Privacy-preserving Statistics

Correlation Measure: Bayesian Networks



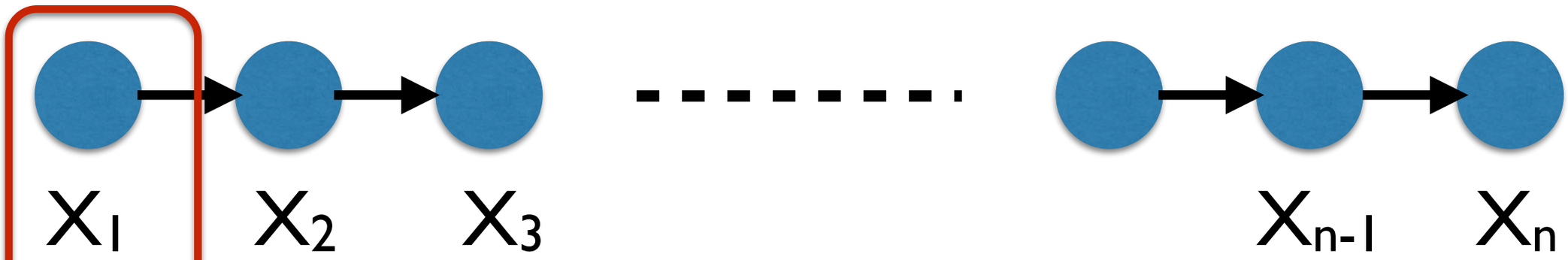
Node: variable

Directed Acyclic Graph

Joint distribution of variables:

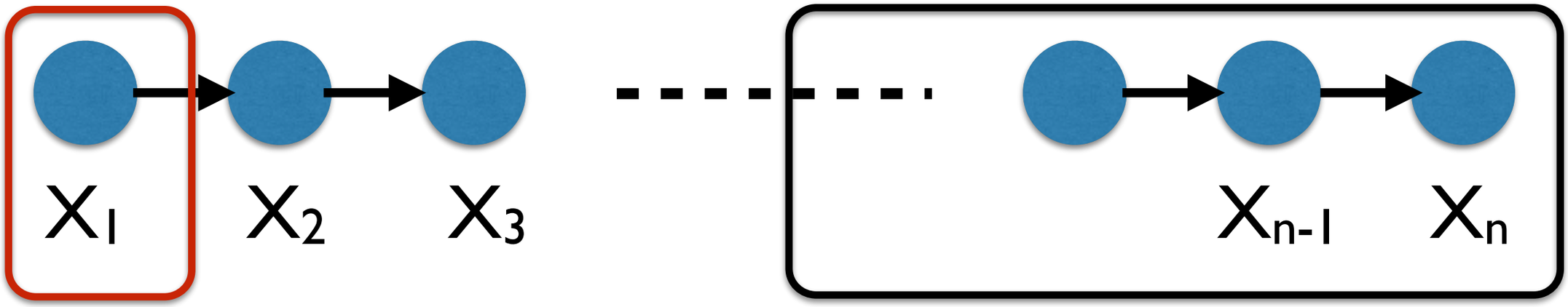
$$\Pr(X_1, X_2, \ldots, X_n) = \prod_i \Pr(X_i | \text{parents}(X_i))$$

Algorithm: Main Idea



Goal: Protect X_1

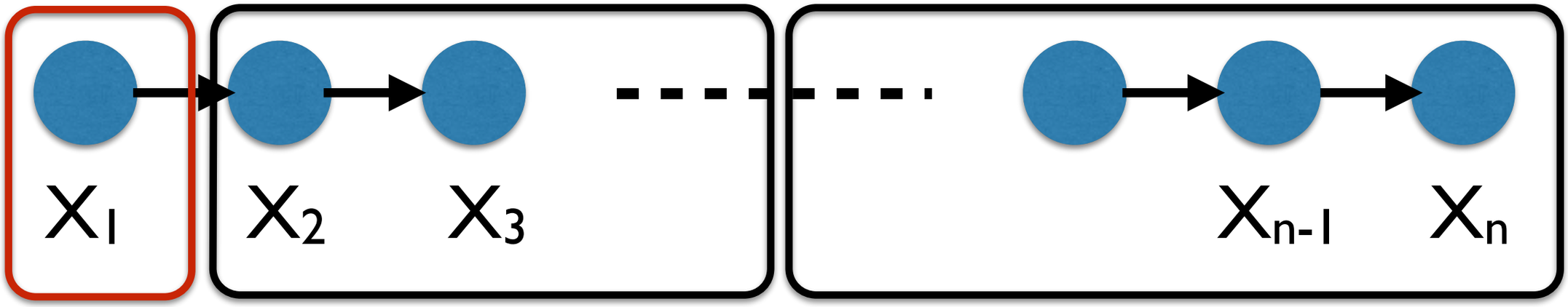
Algorithm: Main Idea



Goal: Protect X_1

Almost independent of X_1

Algorithm: Main Idea



Goal: Protect X_1

Almost independent of X_1

Add noise to hide local terms

+

Small correction for rest

Defining "Almost Independence"

Max-influence of X_i on a set of nodes X_R:

$$e(X_R|X_i) = \max_{a,b} \sup_{\theta \in \Theta} \max_{x_R} \log \frac{\Pr(X_R = x_R | X_i = a, \theta)}{\Pr(X_R = x_R | X_i = b, \theta)}$$

Low $e(X_R|X_i)$ means X_R is almost independent of X_i

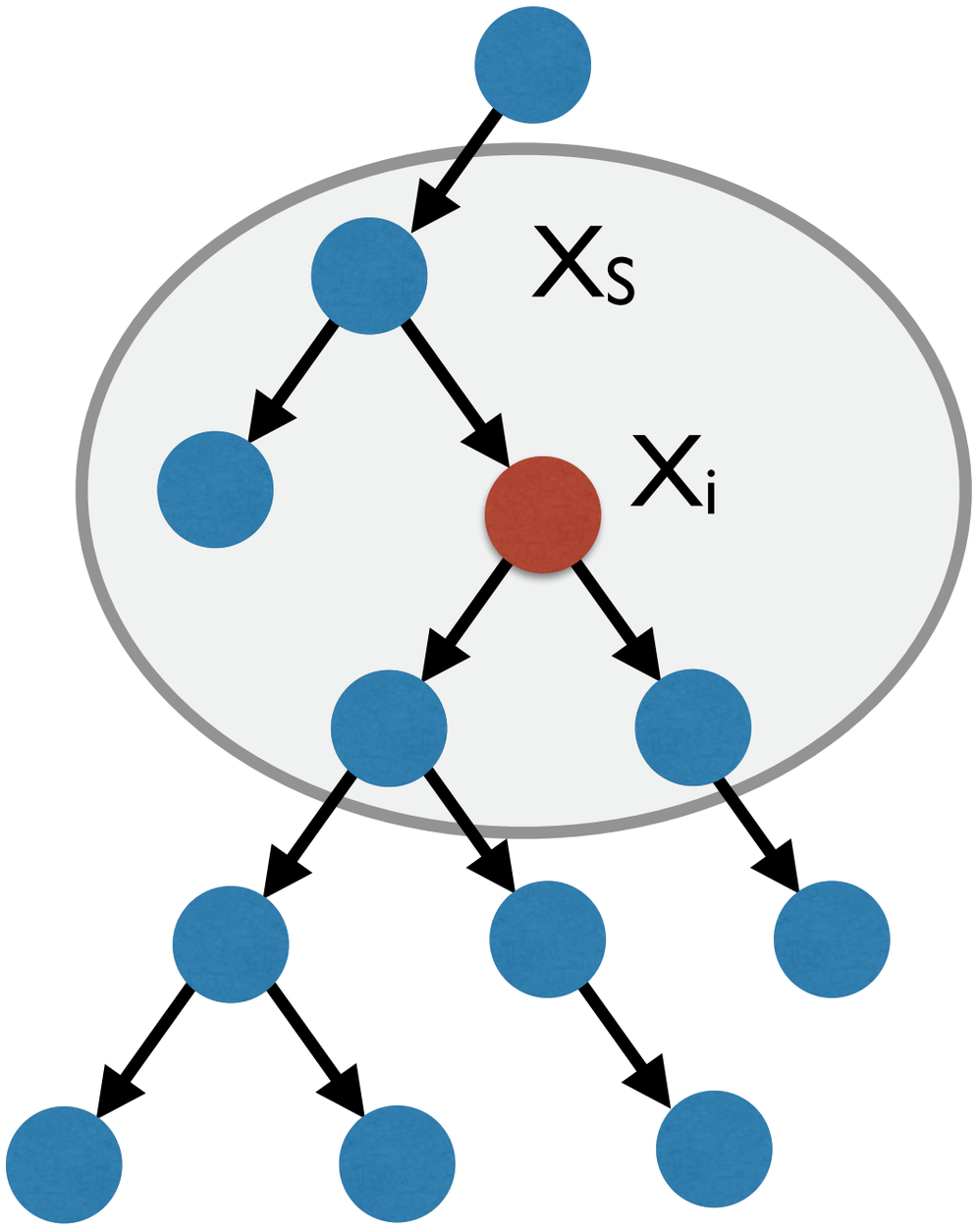
Fact: To protect X_i, correction term needed for X_R is $\exp(e(X_R|X_i))$

How to find large “almost independent” sets

Brute force search is expensive

Use structural properties of the Bayesian network

Markov Blanket

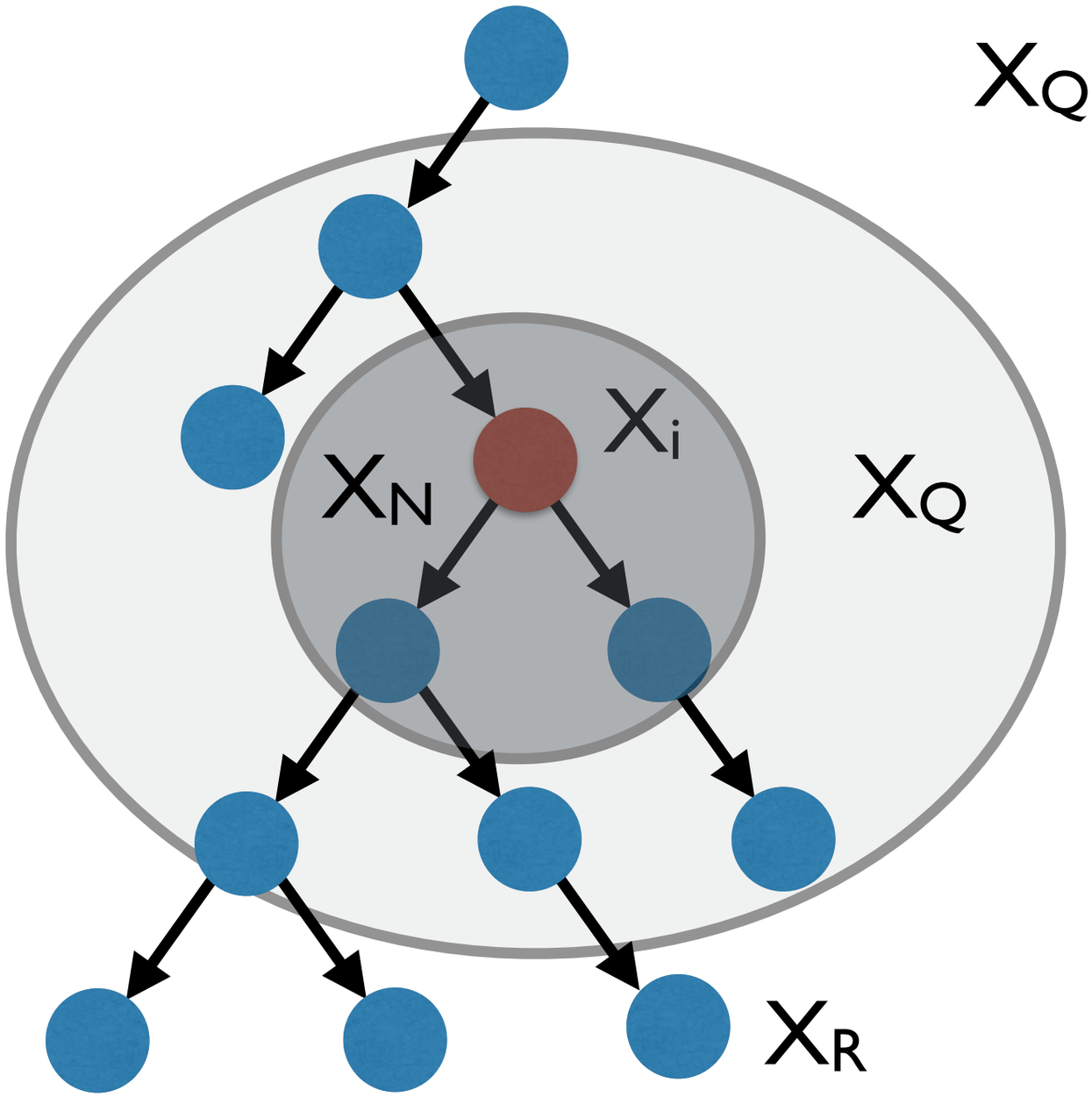


Markov Blanket(X_i) =

Set of nodes X_S s.t

X_i **independent of** $X \setminus (X_i \cup X_S)$ given X_S

Define: Markov Quilt



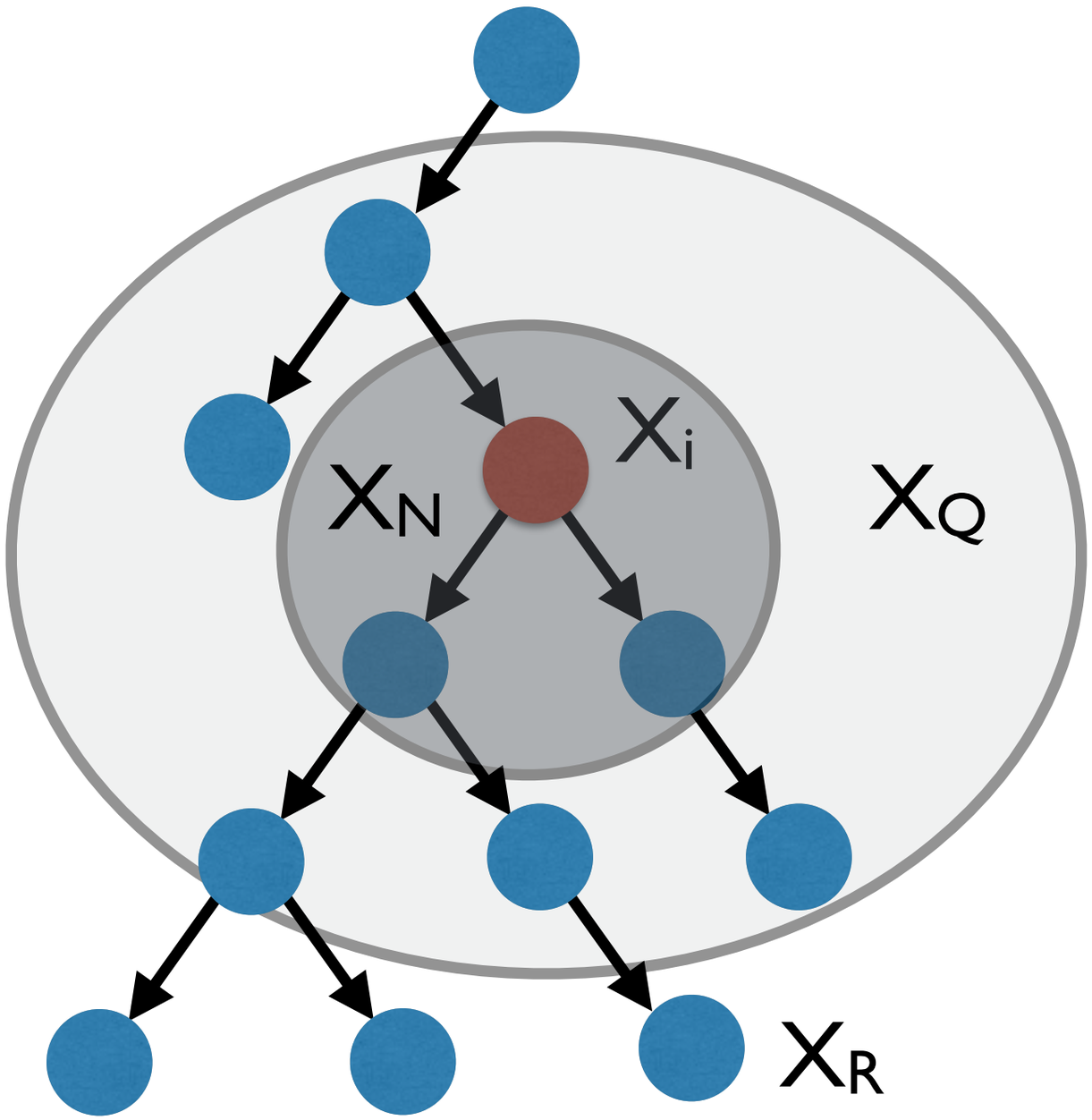
X_Q is a Markov Quilt for X_i if:

1. Deleting X_Q breaks graph into X_N and X_R

2. X_i lies in X_N

3. X_R is independent of X_i given X_Q

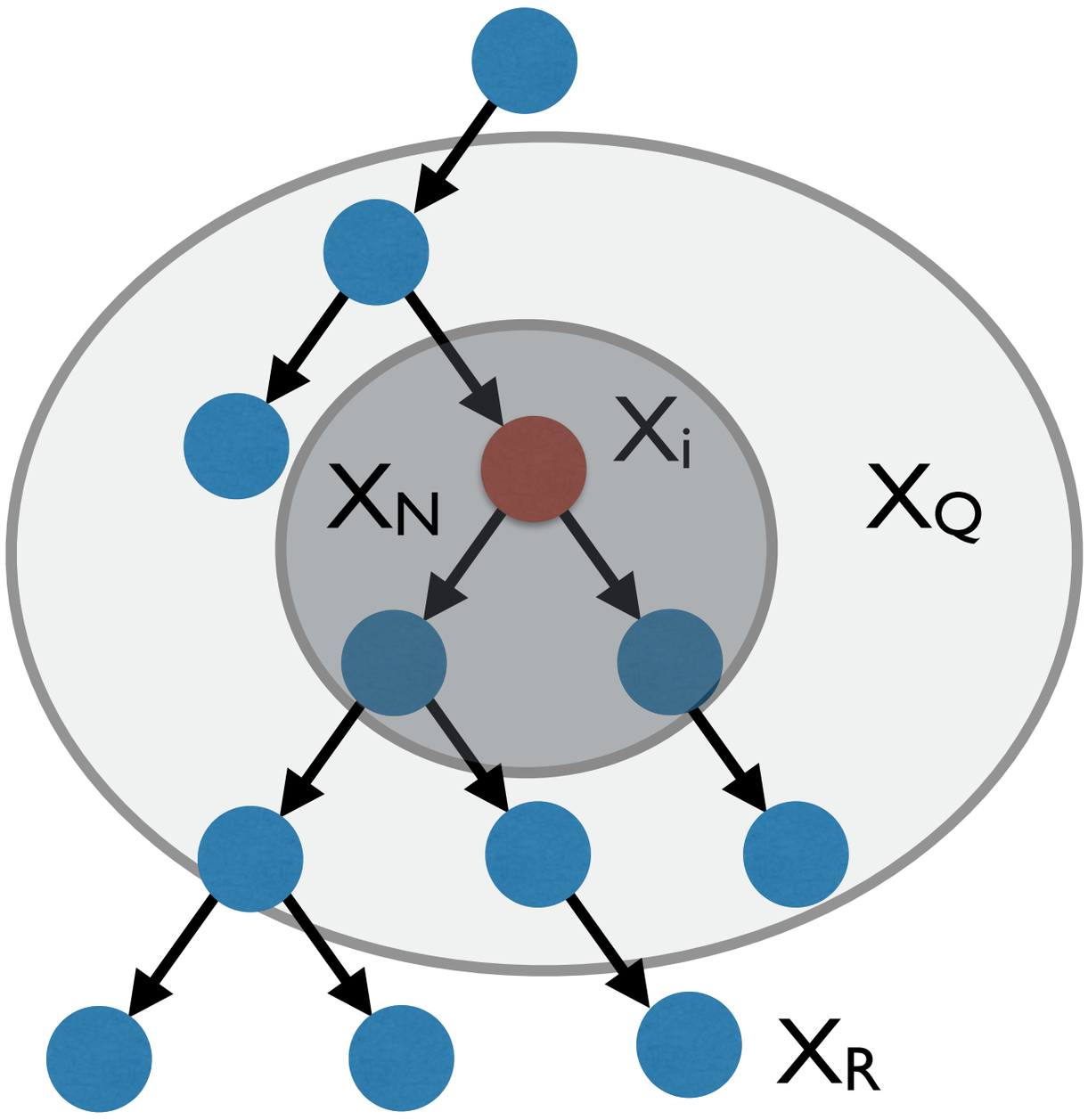
Markov Quilt



Define:

$$\text{Score}(X_Q) = \frac{card(X_N)}{\alpha - e(X_Q|X_i)}$$

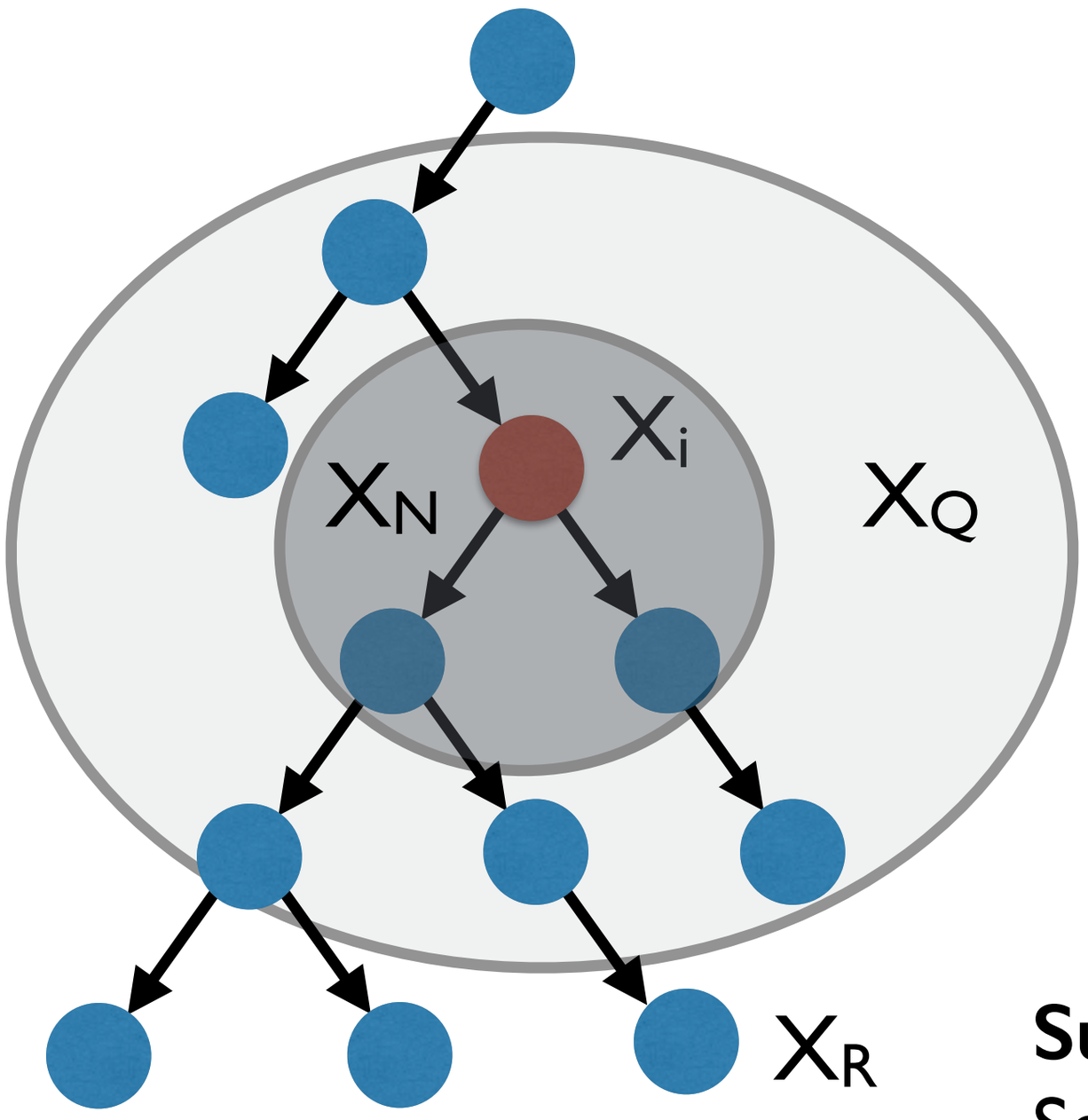
Markov Quilt



$$\text{Score}(X_Q) = \frac{card(X_N)}{\alpha - e(X_Q|X_i)}$$

Correction due to X_Q

Markov Quilt



Stdev of Noise added due to X_N

$$\text{Score}(X_Q) = \frac{card(X_N)}{\alpha - e(X_Q|X_i)}$$

Correction due to X_Q

Sufficient to add noise with stdev Score(X_Q) to protect X_i

The Markov Quilt Mechanism

For each X_i

Find the Markov Quilt X_Q for X_i with minimum score s_i

Output F(D) + ($\max_i s_i$) Z where $Z \sim Lap(1)$

The Markov Quilt Mechanism

For each X_i

Find the Markov Quilt X_Q for X_i with minimum score s_i

Output F(D) + ($\max_i s_i$) Z where $Z \sim Lap(1)$

Theorem: This preserves α-Pufferfish privacy

The Markov Quilt Mechanism

For each X_i

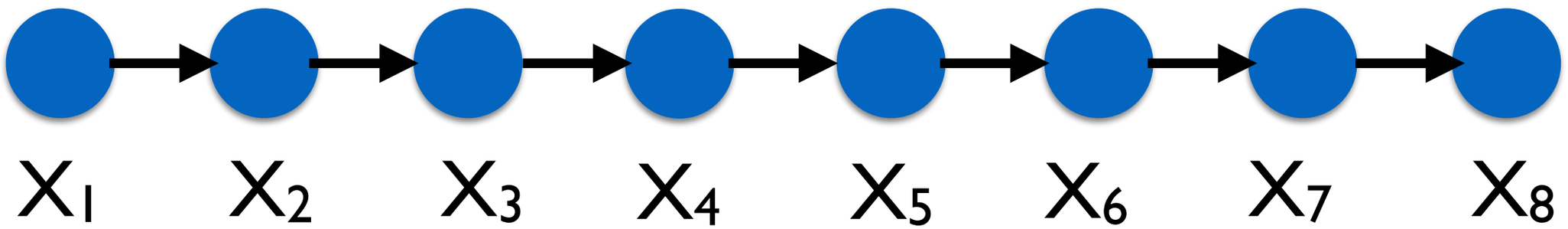
Find the Markov Quilt X_Q for X_i with minimum score s_i

Output F(D) + ($\max_i s_i$) Z where $Z \sim Lap(1)$

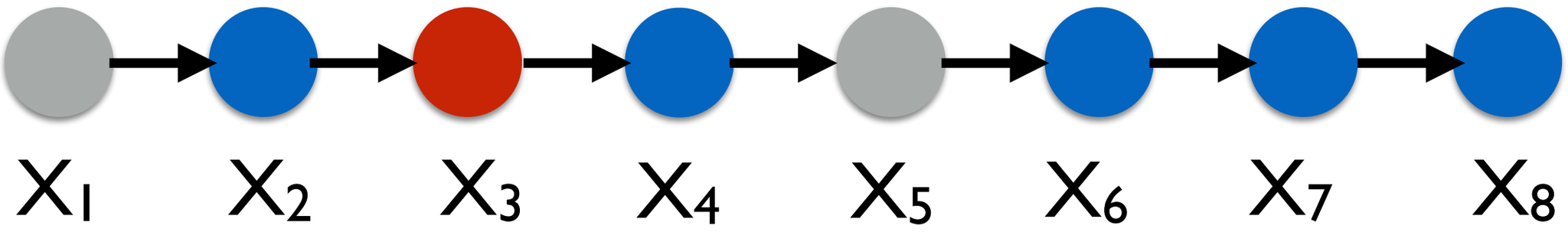
Advantage: Poly-time in special cases.

Example: Activity Monitoring

Bayesian Network: Markov Chain

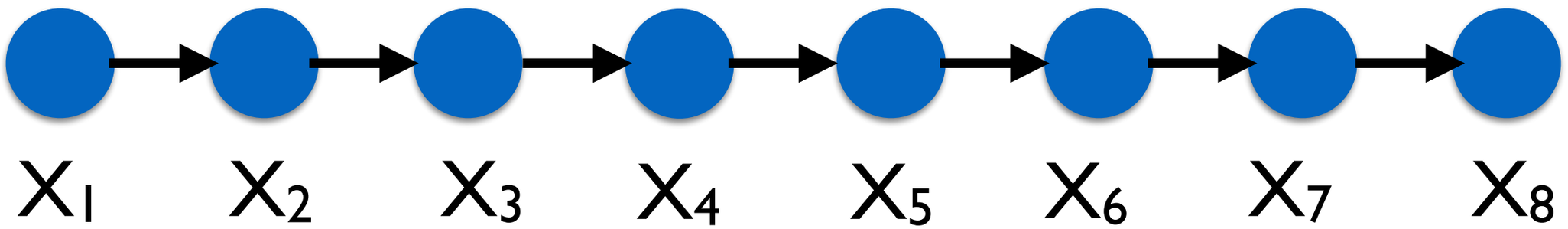


(Minimal) Markov Quilts: for X_i have form $\{X_{i-a}, X_{i+b}\}$

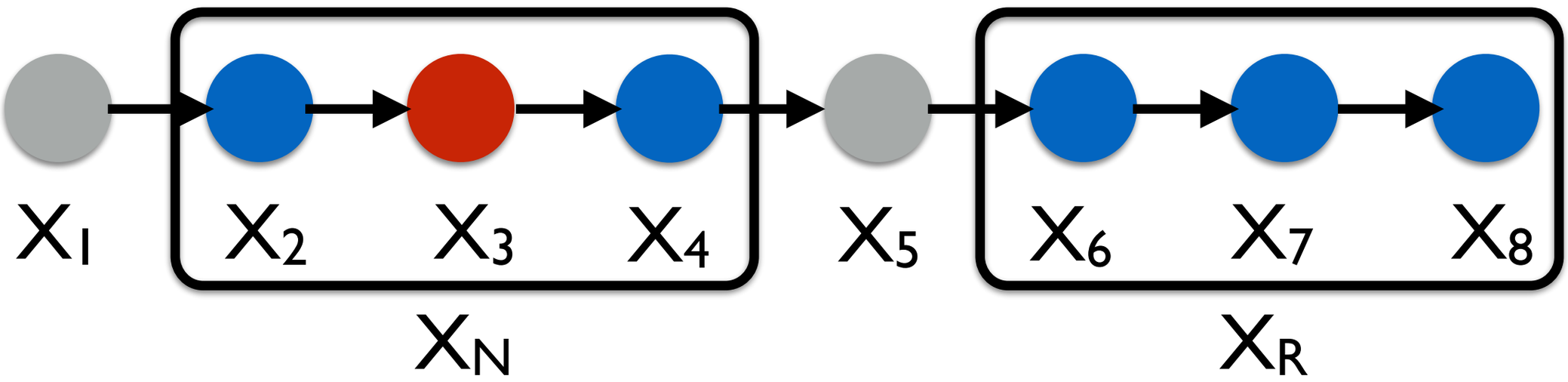


Example: Activity Monitoring

Bayesian Network: Markov Chain



(Minimal) Markov Quilts: for X_i have form $\{X_{i-a}, X_{i+b}\}$



Example: Activity Monitoring: Scores

$\mathcal{X}$: set of states

P_θ : transition matrix describing each $\theta \in \Theta$

Example: Activity Monitoring: Scores

$\mathcal{X}$: set of states

P_θ : transition matrix describing each $\theta \in \Theta$

Under some assumptions, relevant parameters are:

$\pi_\Theta = \min_{x \in \mathcal{X}, \theta \in \Theta} \pi_\theta(x)$ (min prob of x under stationary distr.)

$g_\Theta = \min_{\theta \in \Theta} \min\{1 - |\lambda| : P_\theta x = \lambda x, \lambda < 1\}$ (min eigengap of any P_θ)

Example: Activity Monitoring: Scores

$\mathcal{X}$: set of states

P_θ : transition matrix describing each $\theta \in \Theta$

Under some assumptions, relevant parameters are:

$\pi_\Theta = \min_{x \in \mathcal{X}, \theta \in \Theta} \pi_\theta(x)$ (min prob of x under stationary distr.)

$g_\Theta = \min_{\theta \in \Theta} \min\{1 - |\lambda| : P_\theta x = \lambda x, \lambda < 1\}$ (min eigengap of any P_θ)

Max-influence of $X_Q = \{X_{i-a}, X_{i+b}\}$ for X_i

$$e(X_Q|X_i) \leq \log\left(\frac{\pi_\Theta + \exp(-g_\Theta b)}{\pi_\Theta - \exp(-g_\Theta b)}\right) + 2\log\left(\frac{\pi_\Theta + \exp(-g_\Theta a)}{\pi_\Theta - \exp(-g_\Theta a)}\right)$$

Score(X_Q) = $\dfrac{a+b-1}{\alpha - e(X_Q|X_i)}$

Example: Activity Monitoring

Algorithm:

For each X_i

Find Markov Quilt $X_Q = \{X_{i-a}, X_{i+b}\}$ with minimum score s_i

Output F(D) + ($\max_i s_i$) Z where $Z \sim Lap(1)$

Running Time: $O(T^3)$ (can be made $O(T^2)$)

Conclusion

Problem:

- privacy preserving classification of iid data
- private statistics for correlated data

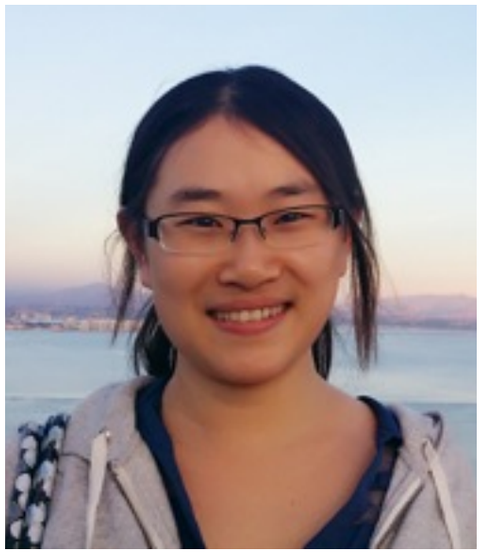
Open Questions:

- better private algorithms for classification
- better models and mechanisms for correlated data

Acknowledgements

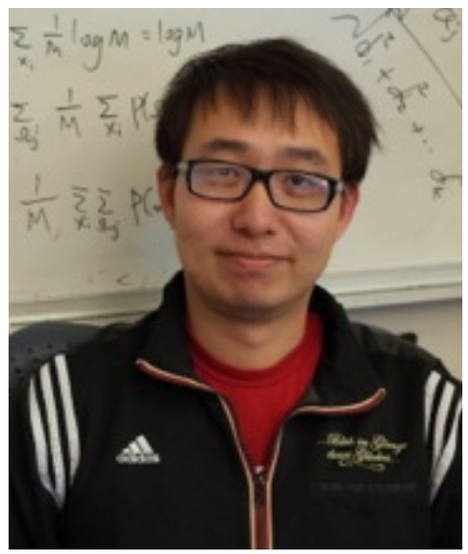
Claire Monteleoni

Anand Sarwate



Shuang Song

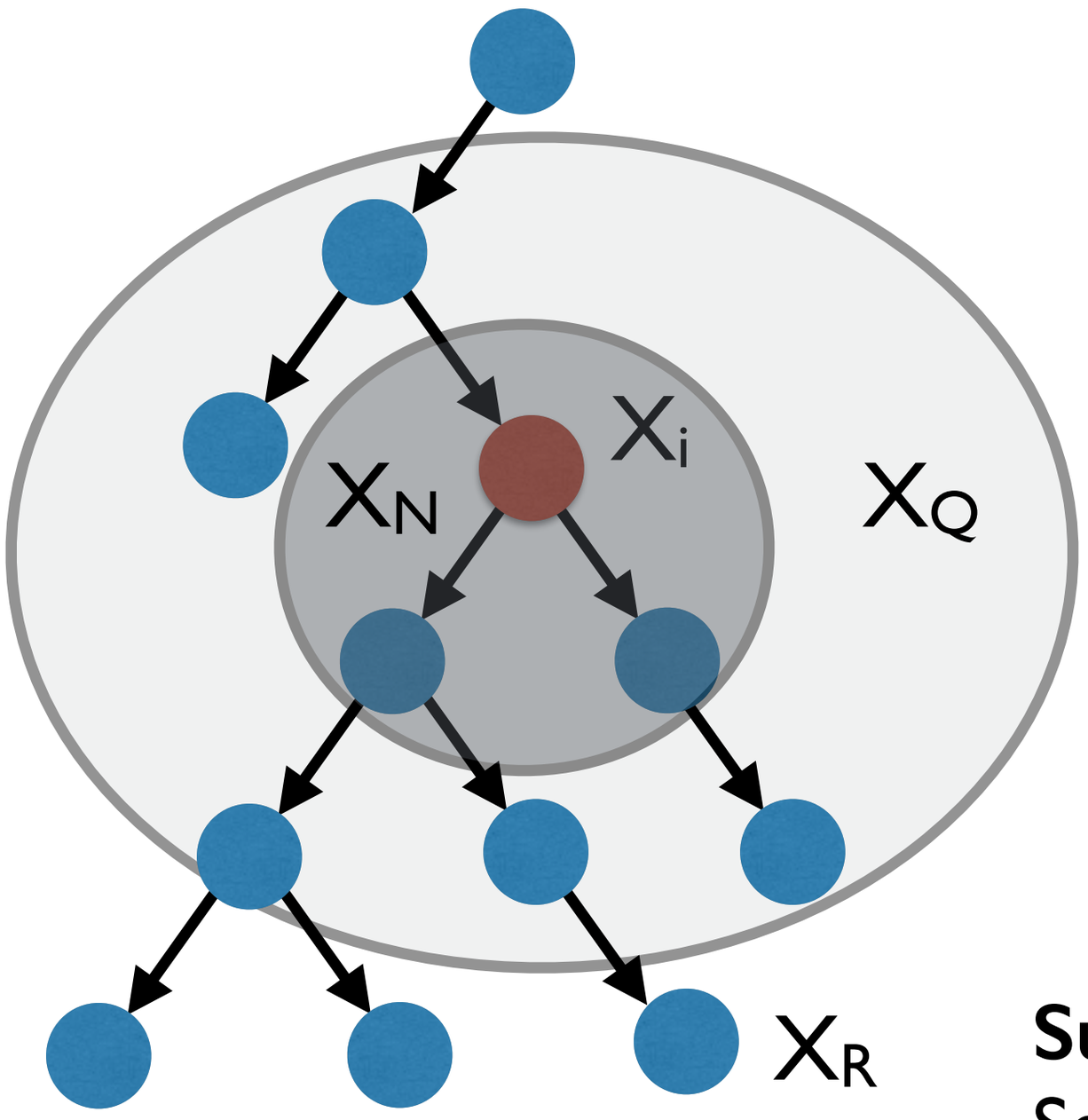
Mani Srivastava



Yizhen Wang

Questions?

Markov Quilt



Stdev of Noise added due to X_N

$$\text{Score}(X_Q) = \frac{card(X_N)}{\epsilon - e(X_Q|X_i)}$$

Correction due to X_Q

Sufficient to add noise with stdev Score(X_Q) to protect X_i