Inference Attacks on Property-Preserving Encrypted Databases

Charles V. Wright
Portland State University
@hackermath

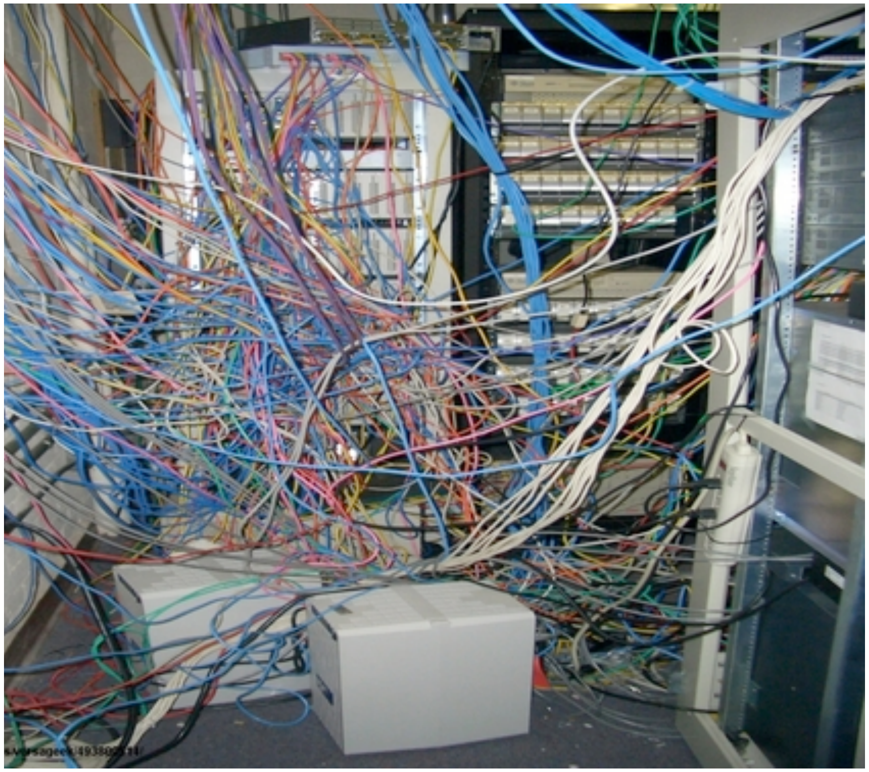
Joint work with
Muhammad Naveed (UIUC/Cornell)
and Seny Kamara (MSR)

“The Cloud”

- Potential for massive cost savings
  - Replace these guys

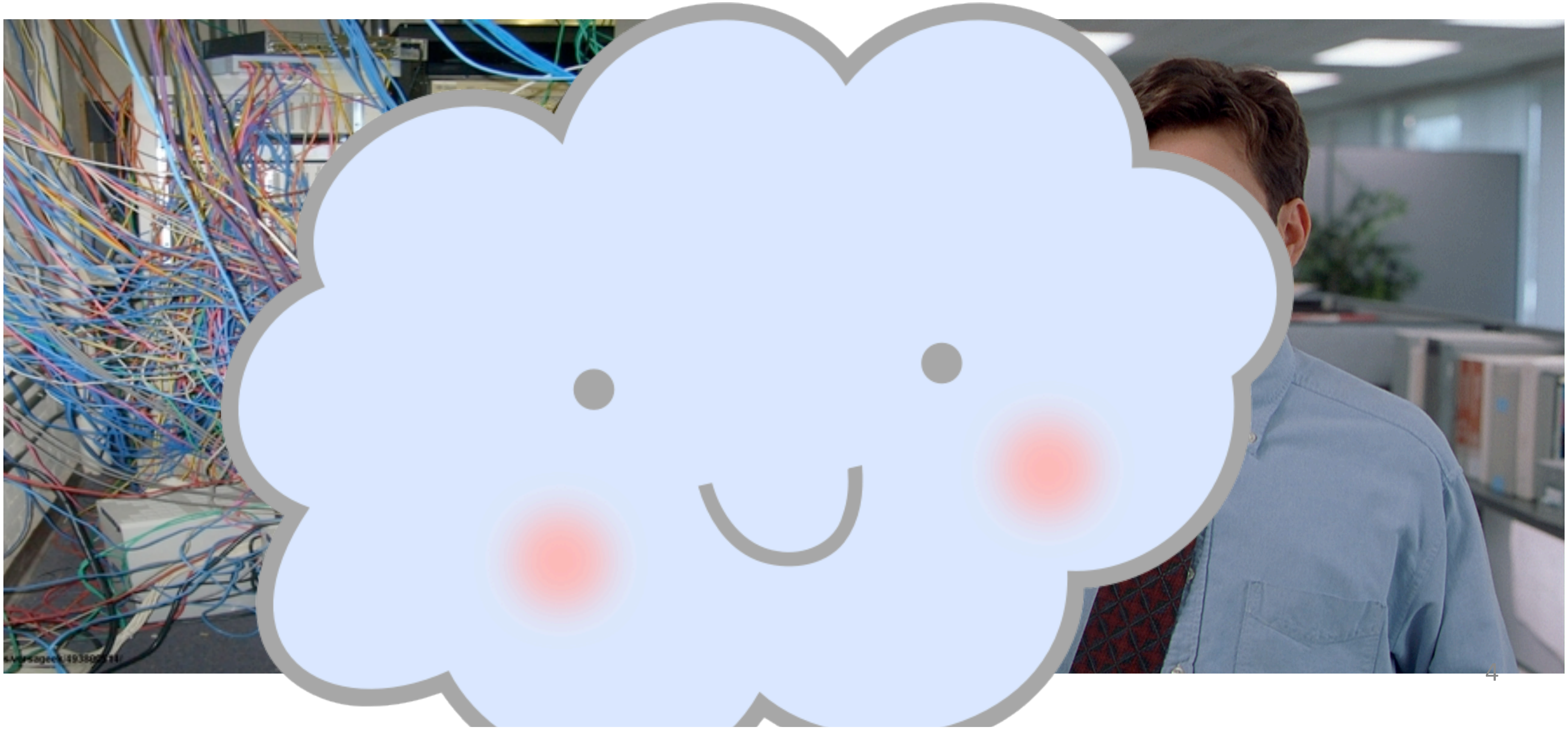
“The Cloud”

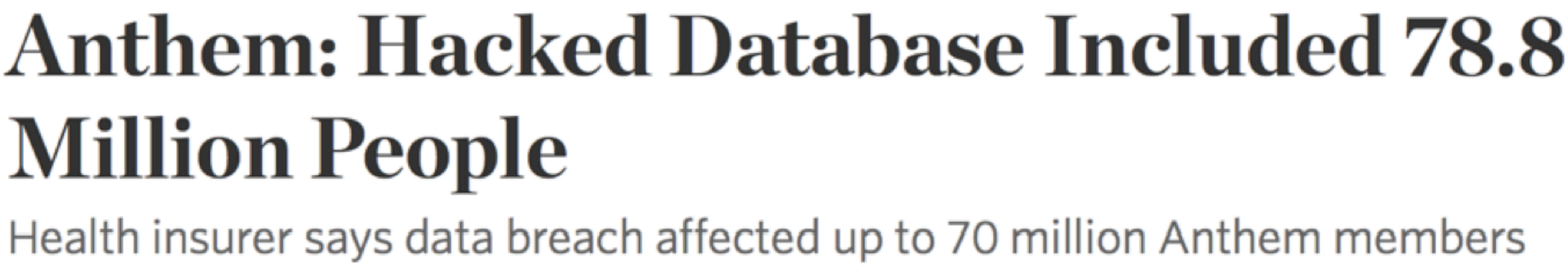
- Potential for massive cost savings
  - Replace this stuff



“The Cloud”

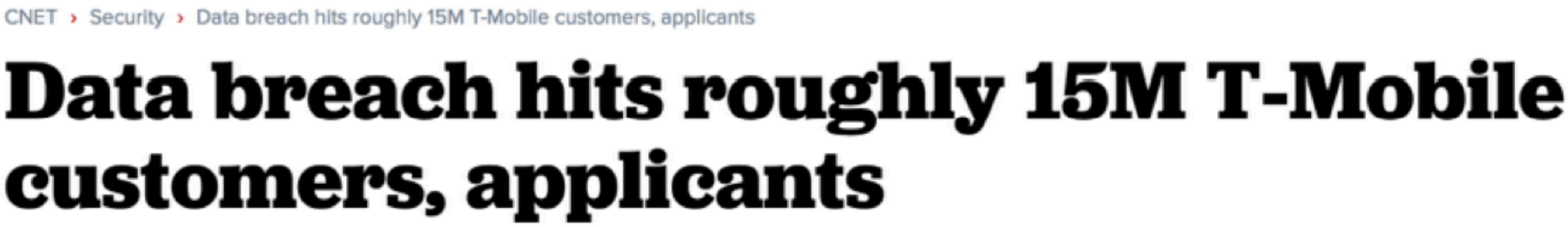
- Potential for massive cost savings
  - With web-based services





Anthem: Hacked Database Included 78.8 Million People

Health insurer says data breach affected up to 70 million Anthem members



CNET › Security › Data breach hits roughly 15M T-Mobile customers, applicants

Data breach hits roughly 15M T-Mobile customers, applicants

JUL 20, 2015 @ 04:18 AM 32,267 VIEWS

Ashley Madison Breach Could Expose Privates Of 37 Million Cheaters

ASHLEY MADISON®
Life is short. Have an affair.®
Get started by telling us your relationship status:
Please Select
See Your Matches »
Over 37,565,000 anonymous members!

Target: 40 million credit cards compromised

Encryption to the rescue! ... Right?

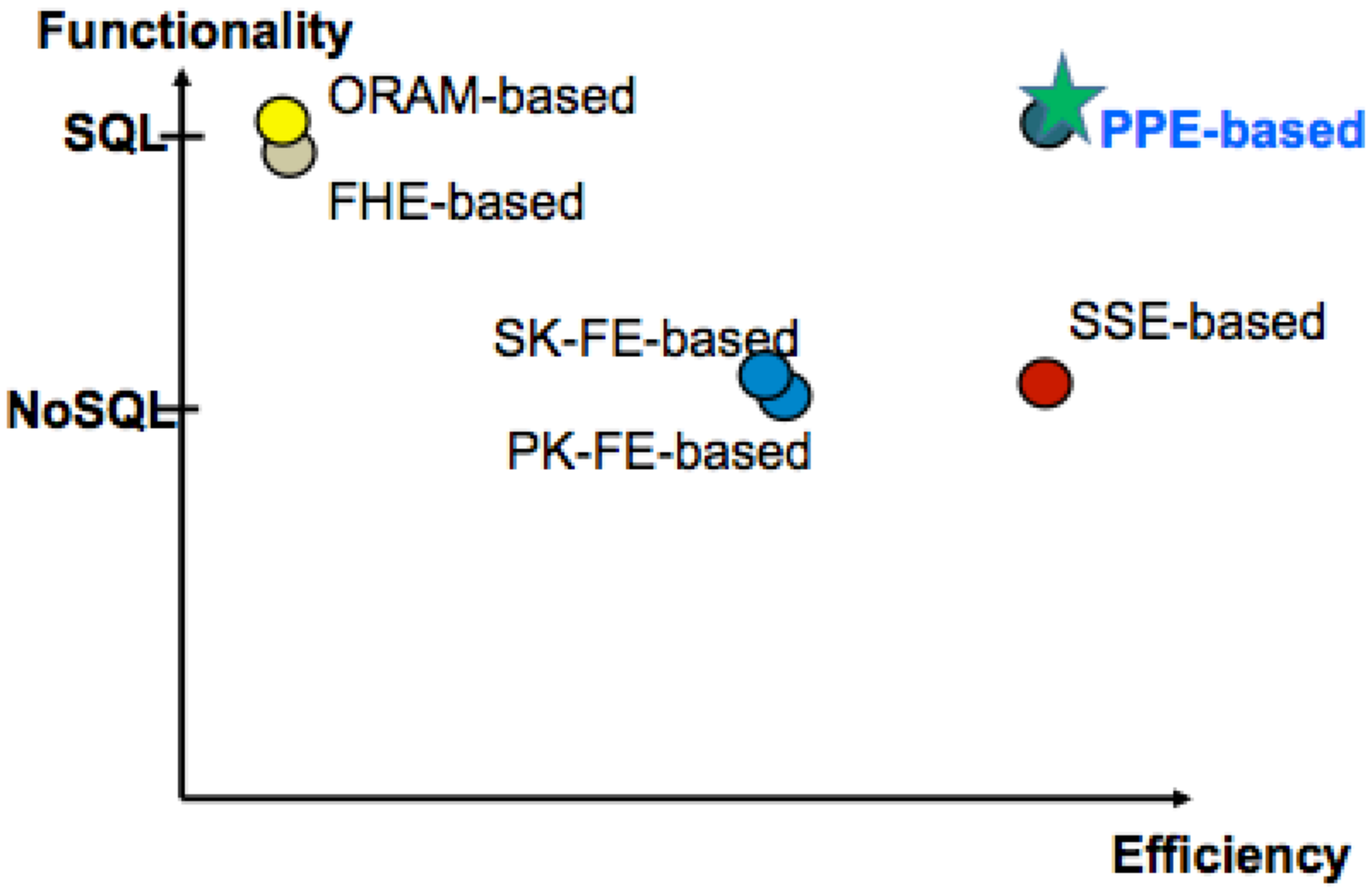
- Not so fast…
  - – Lose search, DBs, IR
  - – How to find your photo among 300PBs?
  - – How to rank results?

SEARCHING ON ENCRYPTED DATA

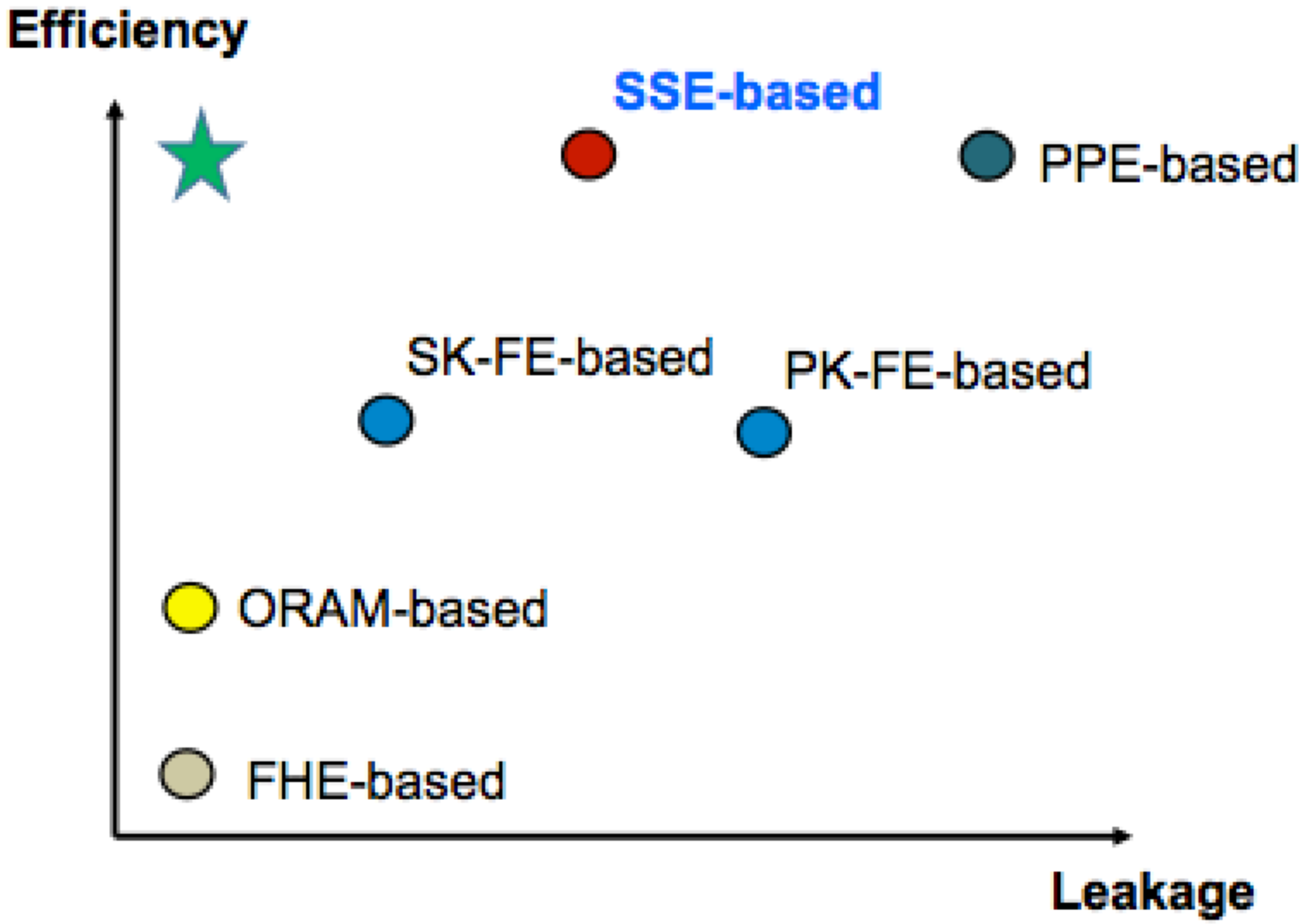
Many Approaches

- Stream ciphers [SWP01]
- Bucketing [HILM02]
- Structured and searchable encryption (StE/SSE) [SWP01,CGKO06,CK10]
- Oblivious RAM (ORAM) [GO96]
- Functional encryption (e.g., PEKS) [BCOP06]
- Multi-party computation (MPC)
- **Property-preserving encryption (PPE)** [AKSX04,BBO06,BCLO09]
- Efficiently Searchable Encryption [HAJSS14, LCSJLB14]
- Fully-homomorphic encryption [G09]

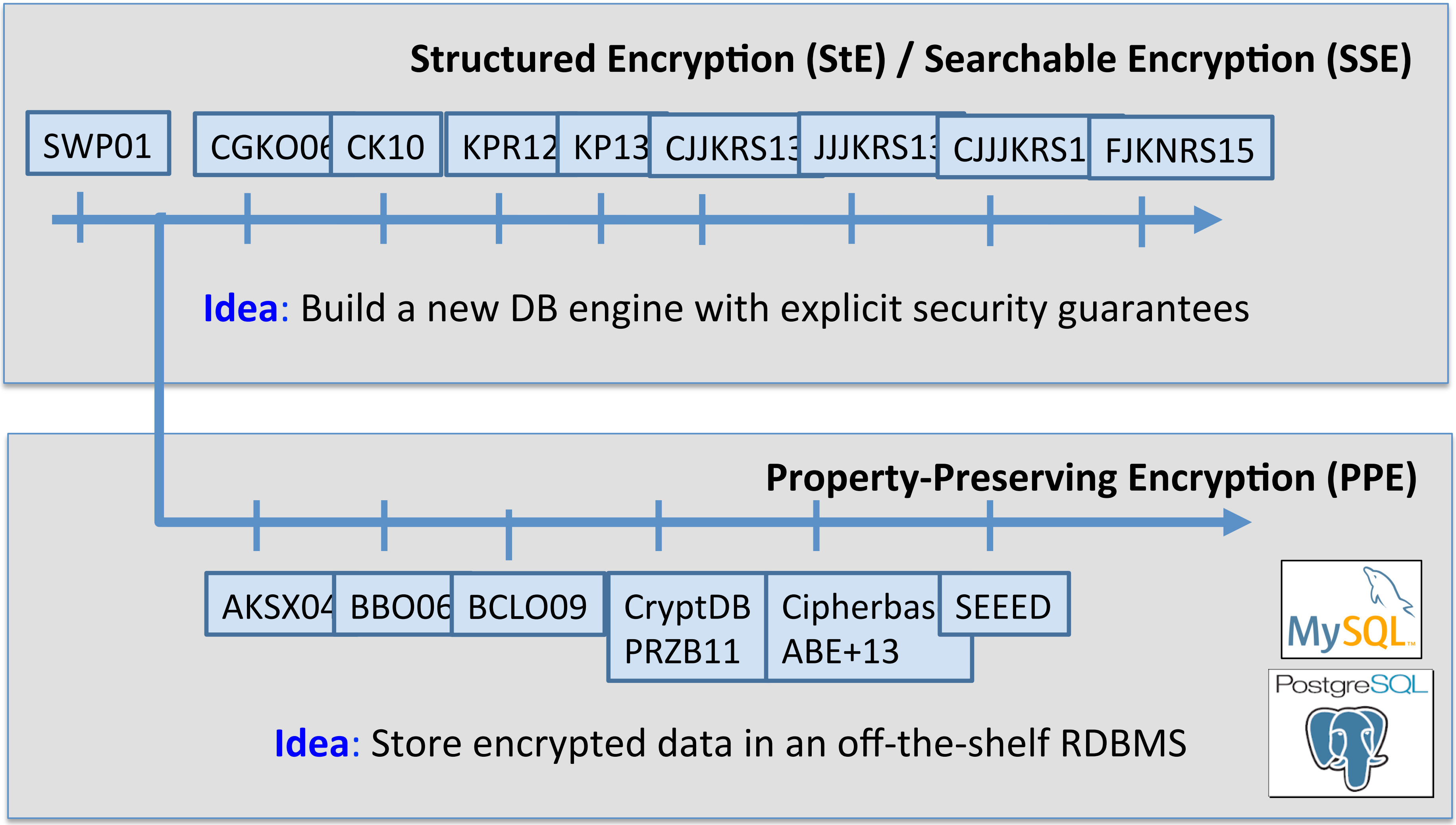
Tradeoffs: Functionality vs Efficiency



Tradeoffs: Efficiency vs Leakage



Two Branches of Research



Structured Encryption (StE) / Searchable Encryption (SSE)

SWP01 CGKO06 CK10 KPR12 KP13 CJJKRS13 JJJKRS13 CJJJKRS1 FJKNRS15

Idea: Build a new DB engine with explicit security guarantees

Property-Preserving Encryption (PPE)

AKSX04 BBO06 BCLO09 CryptDB PRZB11 Cipherbas ABE+13 SEEED

Idea: Store encrypted data in an off-the-shelf RDBMS

Property-Preserving Encryption

Standard Encryption

Age	Age
19	LKGM8EUnGd
32	kt6gUXGWgL
22	TRxZDzVYjV
22	IgDwwF64cl

Deterministic

Age	Age
19	LKGM8EUnGd
32	kt6gUXGWgL
22	**TRxZDzVYjV**
22	**TRxZDzVYjV**

Order-Preserving

Age	Age
19	7399
32	20306
22	**10416**
22	**10416**

- Encryption schemes that reveal/leak properties of plaintext
  - Weaker than standard encryption
  - Enable operations on encrypted data *without homomorphic operations*
  - Deterministic encryption leaks equality
  - Order-preserving encryption (OPE) leaks order

PPE-Based EDBs

- CryptDB [PRZB11]
  - Handles large subset of SQL
  - Very efficient (14-26% overhead)
- Cipherbase [ABEKKRV13]
  - Handles all of SQL
  - PPE + trusted hardware
- SEEED [GHHKKSST14]
  - Handles subset of SQL
  - CryptDB integrated into SAP's HANA DB
- Software from SAP, Google, Microsoft, and others

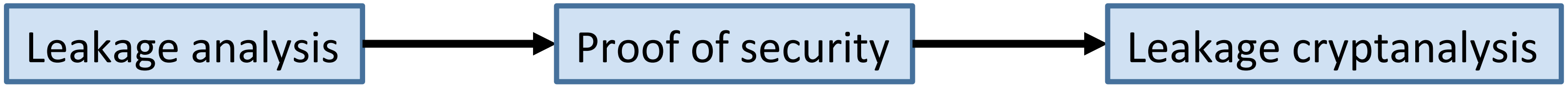
PPE-Based EDBs

- Some PPE-capable systems also include more secure, more expensive modes as alternatives
  - CipherBase – special hardware
  - CryptDB – client-side processing, etc.

- **Cryptanalysis helps users know when to fall back on these alternatives**

Evaluating Security

[Curtmola-Garay-Kamara-Ostrovsky06, Chase-Kamara10, Islam-Kuzu-Kantarcioglu12]



Leakage analysis → Proof of security → Leakage cryptanalysis

- Leakage analysis: what is being leaked?
- Proof: prove that solution leaks no more
- Cryptanalysis: can we exploit the leakage?

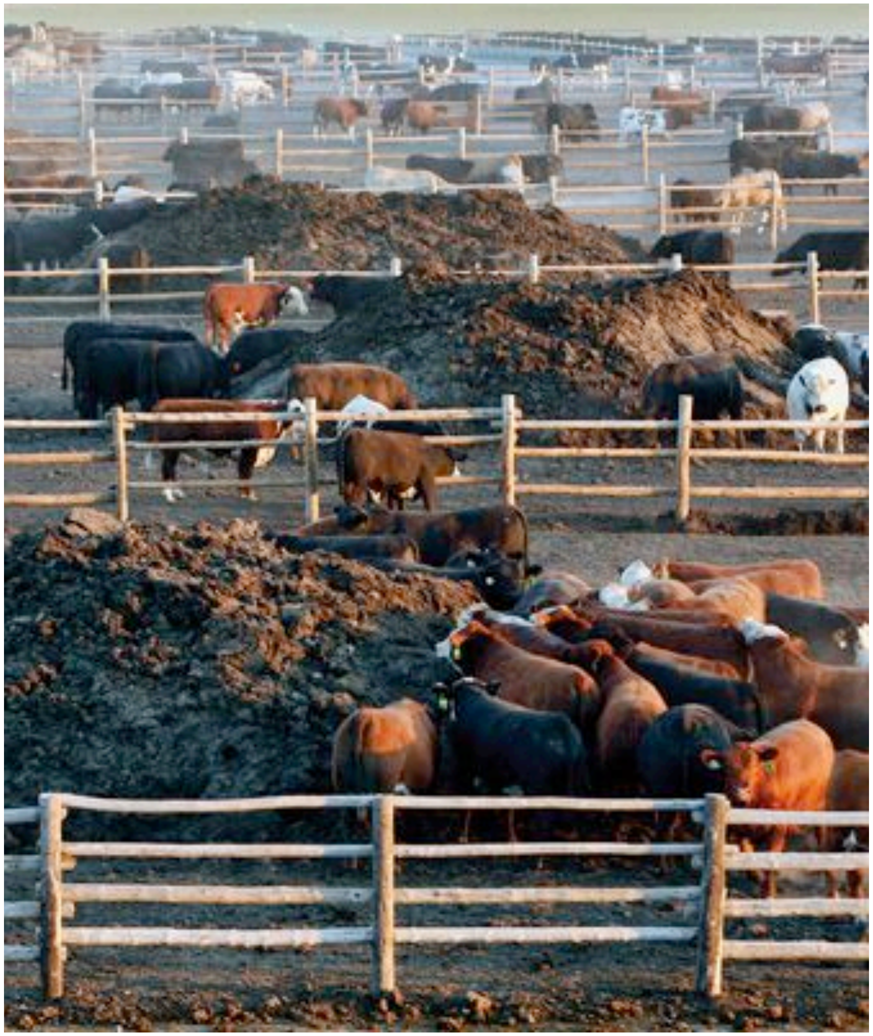
Understanding Leakage of PPE

- Maybe it's not so bad…?
- Previous analyses proved security of DTE and OPE under ideal conditions
  - High min-entropy [BBO07]
  - Uniform random data [BCLO09]
- These works are a great start, but …

What Happens in the Real World?

- Real cows are not spherical or cute

- Real data tends to be
  - Non-uniform
  - Low entropy

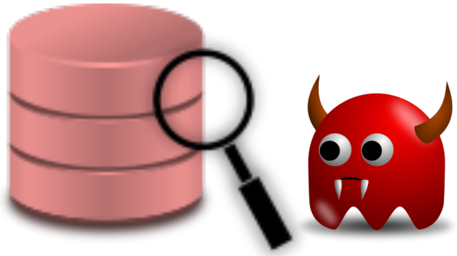


INFERENCE ATTACKS

Inference Attacks

- Adversary has some source of auxiliary information with stats similar to those of the plaintext

- Adversary observes the ciphertext, and collects the same stats



- He puts the two together to make good guesses about the plaintext

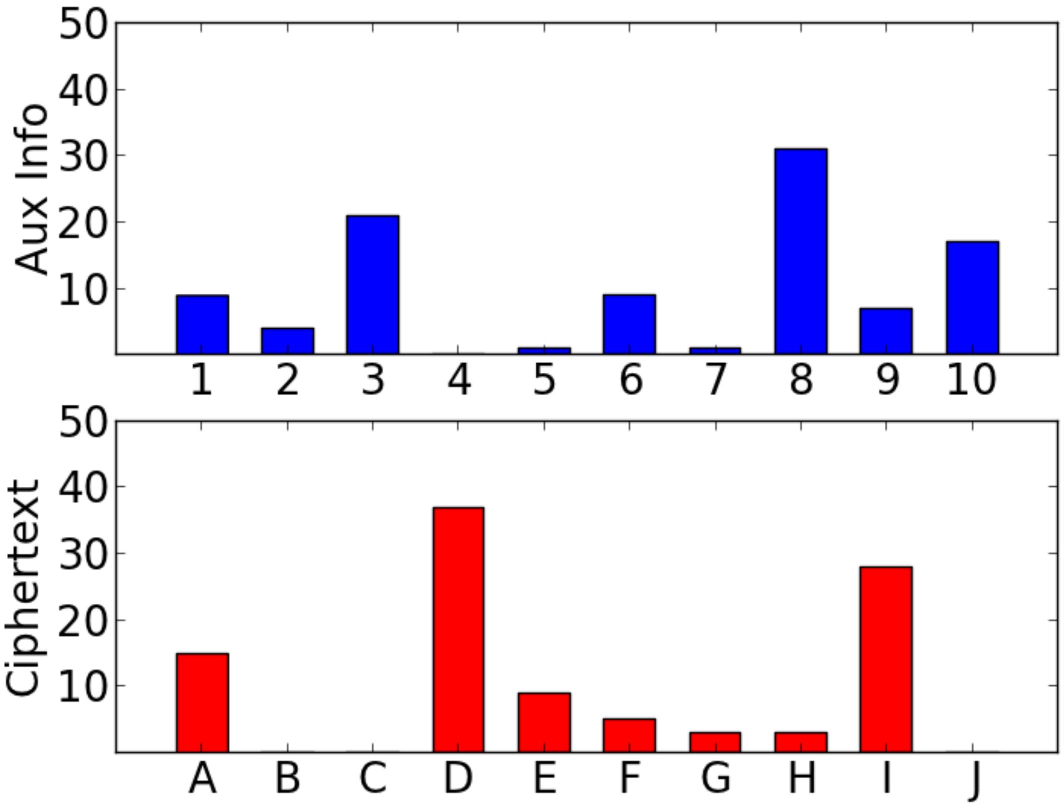
Inference Attacks on PPE

- Two well-known attacks
  - Frequency Analysis [Al-Kindi, 9th century]
  - Sorting Attack [folklore]

- Two new attacks based on combinatorial optimization [NKW15]
  - Lp-Optimization
  - Cumulative Attack

Inference Attacks on Deterministic Encryption

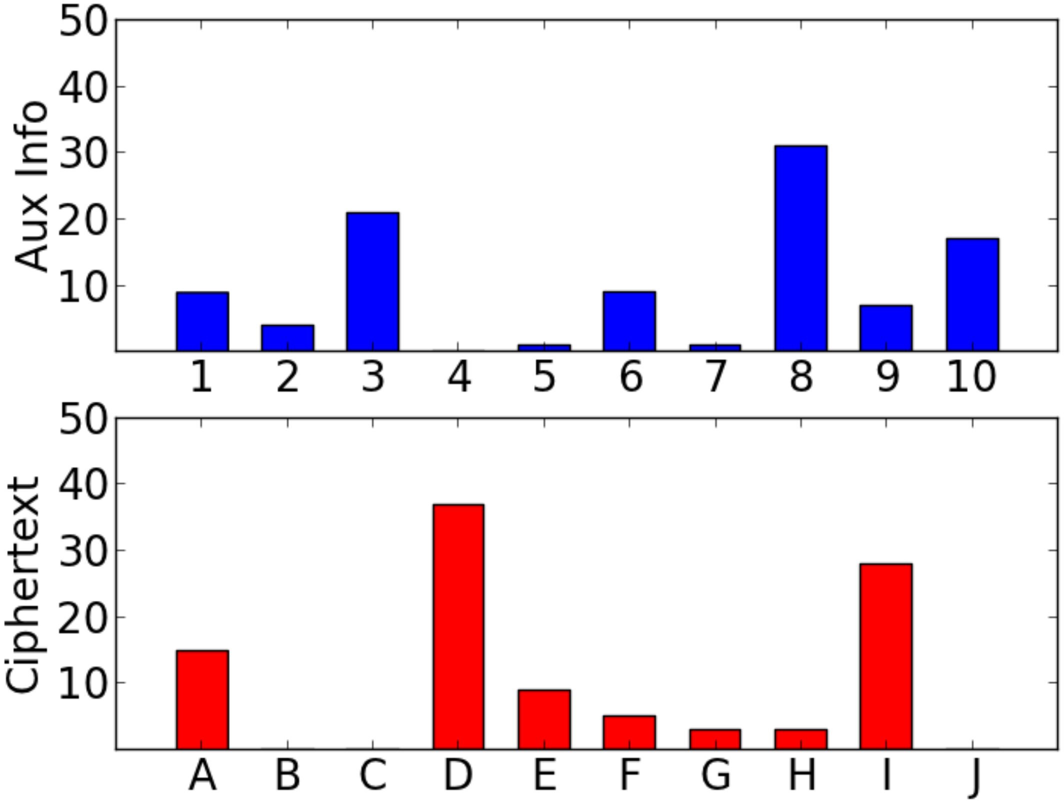
- DTE reveals frequency of the plaintexts
  - ie, the histogram
- Very much like a substitution cipher
  - Think *Intro to Crypto* homework



Manual Cryptanalysis

aka *Just Eyeball It*

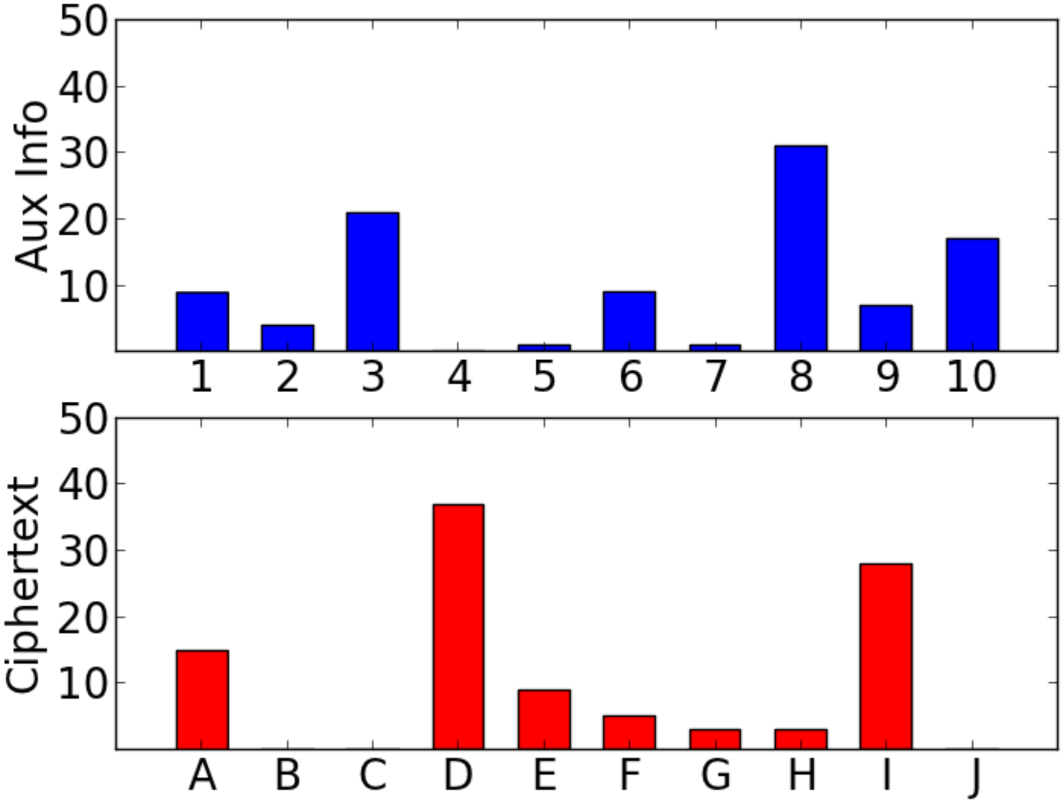
- Looks like
  - 8 = D or maybe I
  - 3 = A or maybe I
  - A = 1 or maybe 10
  - …



- This works OK for *Intro to Crypto* homework

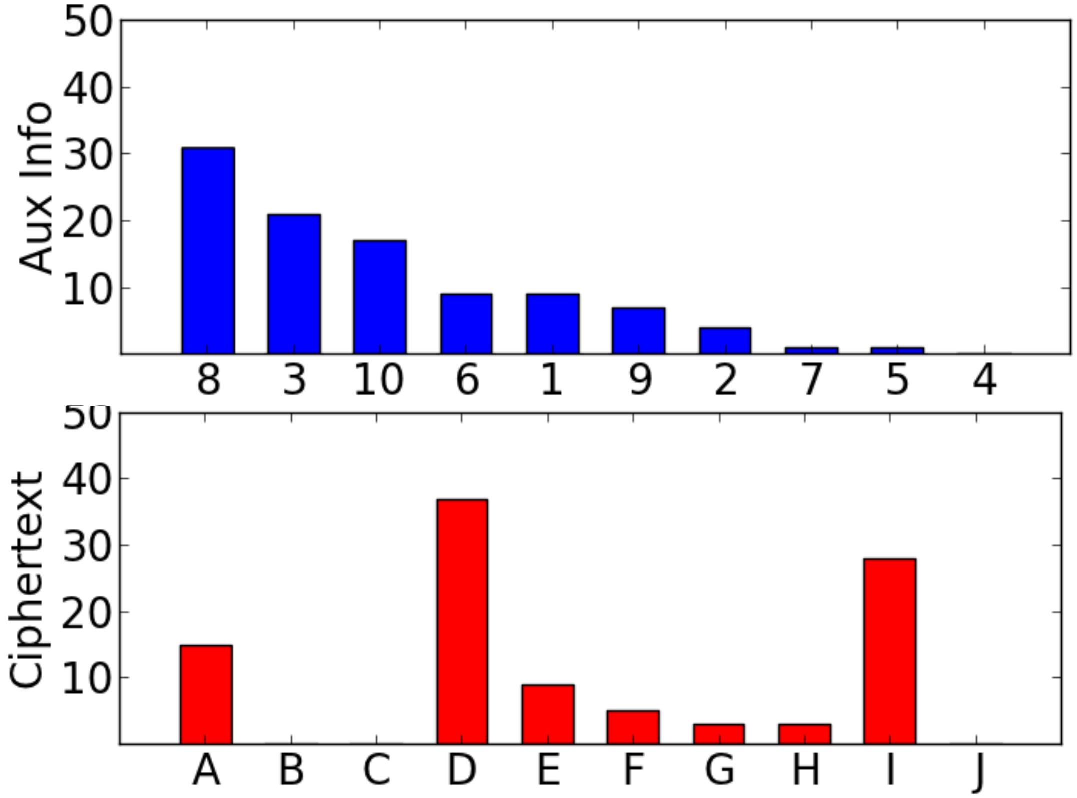
- In the real world, we need an algorithm!

Frequency Analysis (Al-Kindi, 9th century AD)



Frequency Analysis (Al-Kindi, 9th century AD)

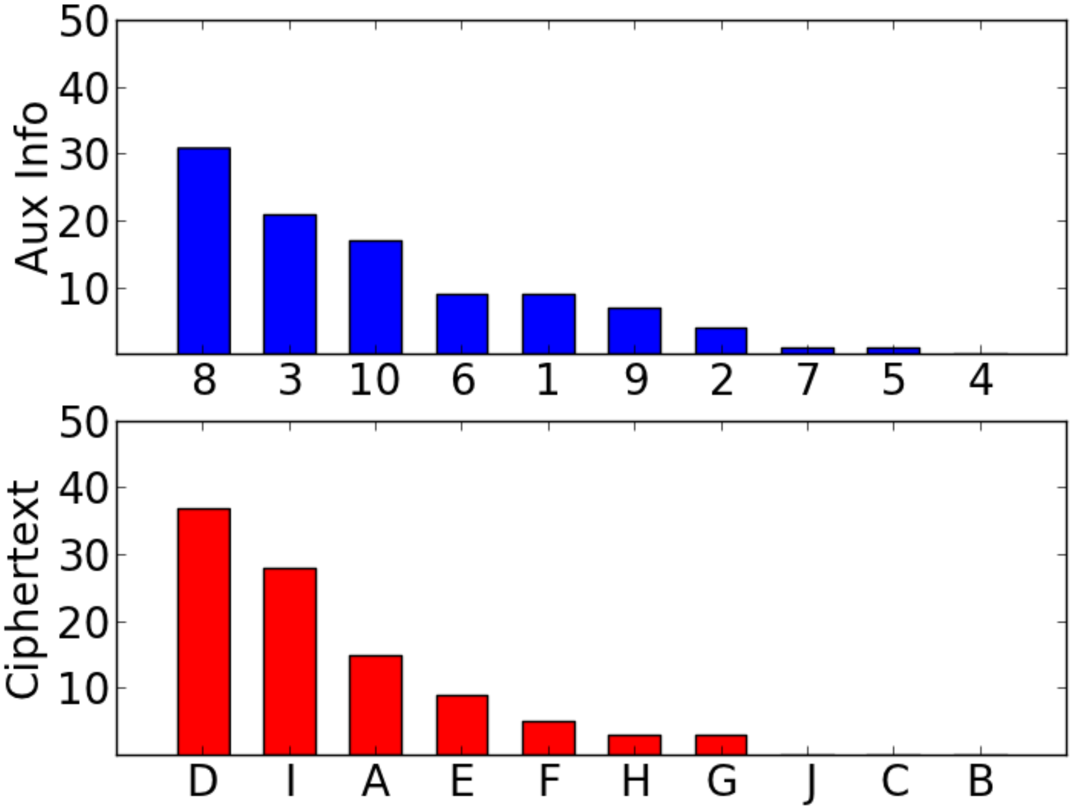
1. Sort plaintexts by aux frequency



Frequency Analysis (Al-Kindi, 9th century AD)

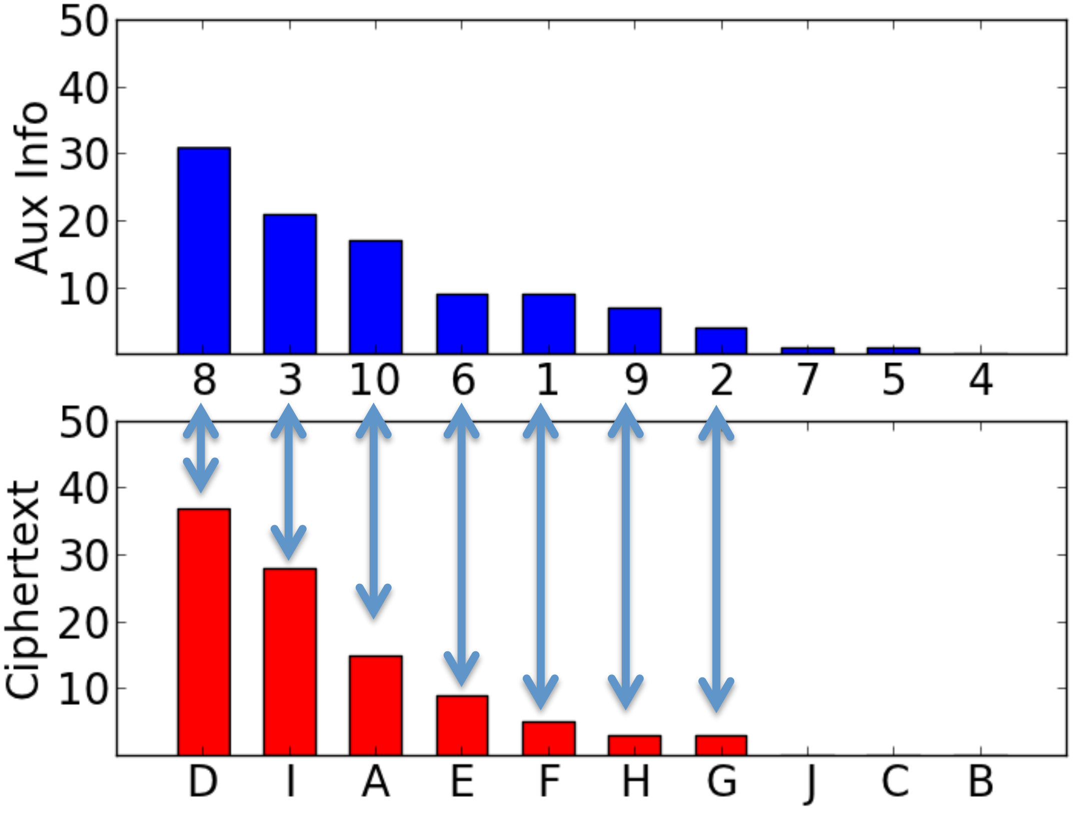
1. Sort plaintexts by aux frequency

2. Sort ciphertexts by frequency



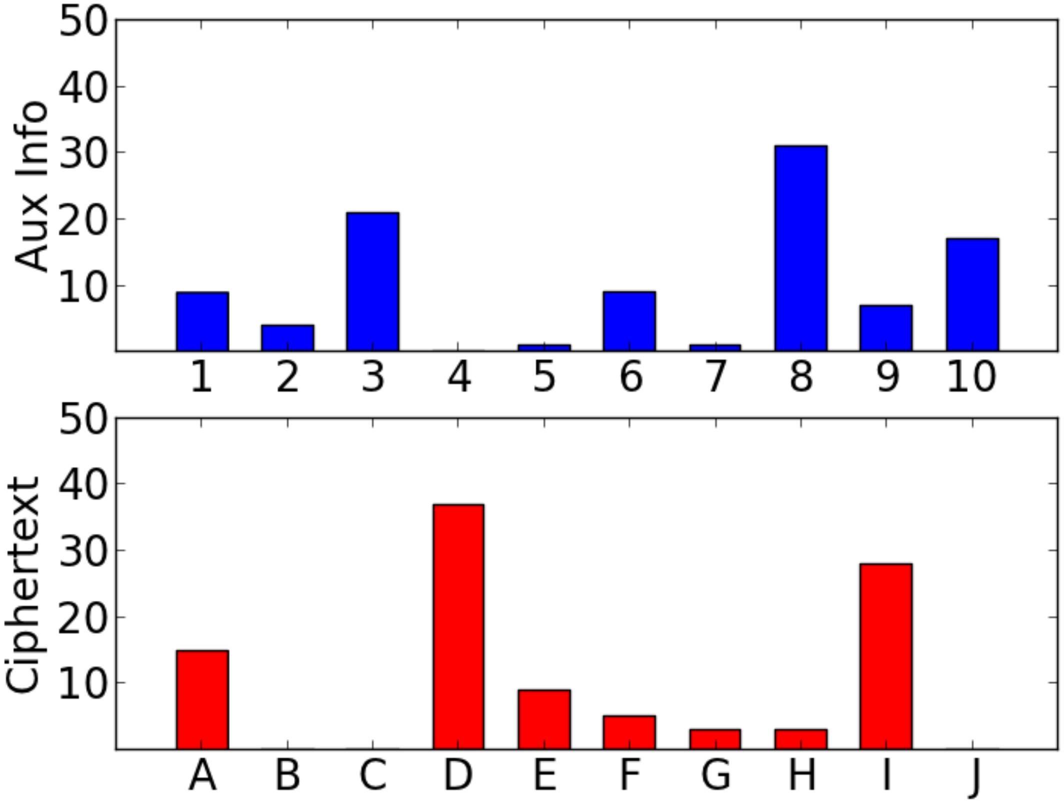
Frequency Analysis (Al-Kindi, 9th century AD)

1. Sort plaintexts by aux frequency
2. Sort ciphertexts by frequency
3. Match them up



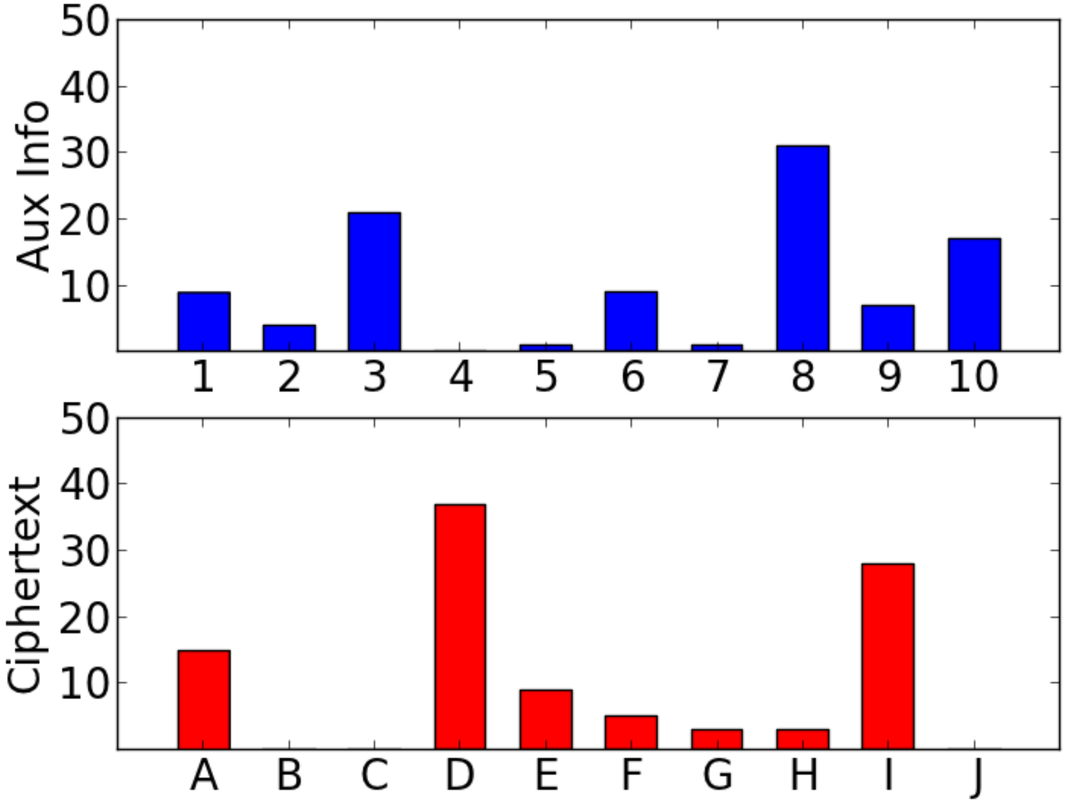
Lp Optimization

- Idea: Find the **best** mapping of plaintexts to ciphertexts based on the histograms



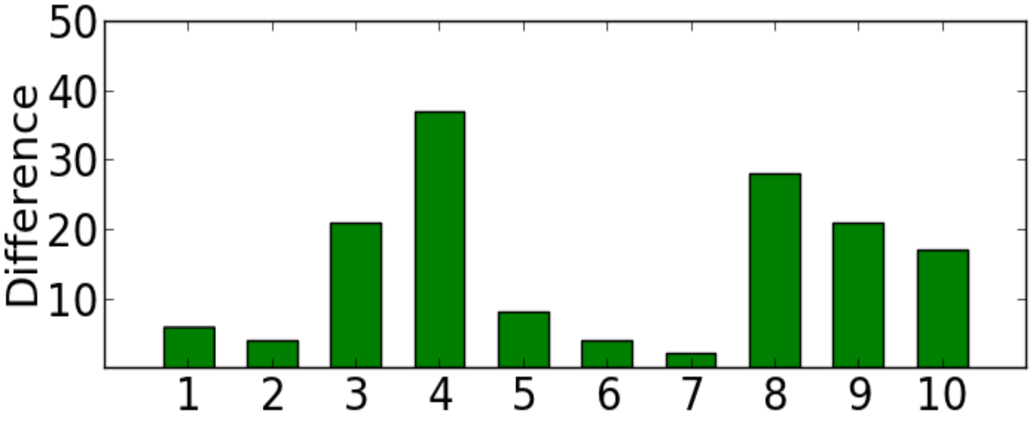
Lp Optimization

- Compute the **difference** in histogram bin heights as a vector



Lp Optimization

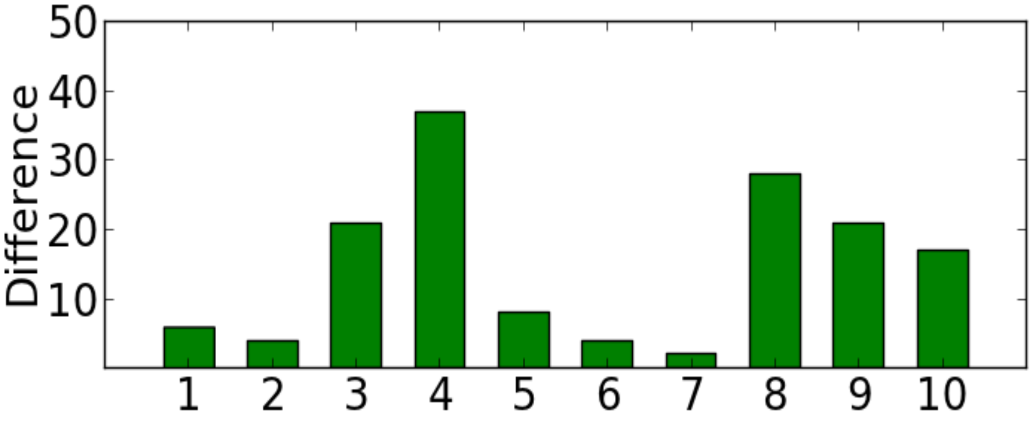
- Compute the **difference** in histogram bin heights as a vector



Lp Optimization

- Compute the difference in histogram bin heights as a vector

- Pick the mapping that minimizes the **Lp norm** of this vector



Lp Optimization

- L1 norm is simply the sum of the differences
  - L1 = 6 + 4 + 7 + 0 + 2 + 0 + 1 + 6 + 2 + 2
- L2 norm is the sum of squared differences
  - L2 = $6^2 + 4^2 + 7^2 + 0^2 + 2^2 + 0^2 + 1^2 + 6^2 + 2^2 + 2^2$
- L3 norm is the sum of cubed differences
- …

Lp Optimization

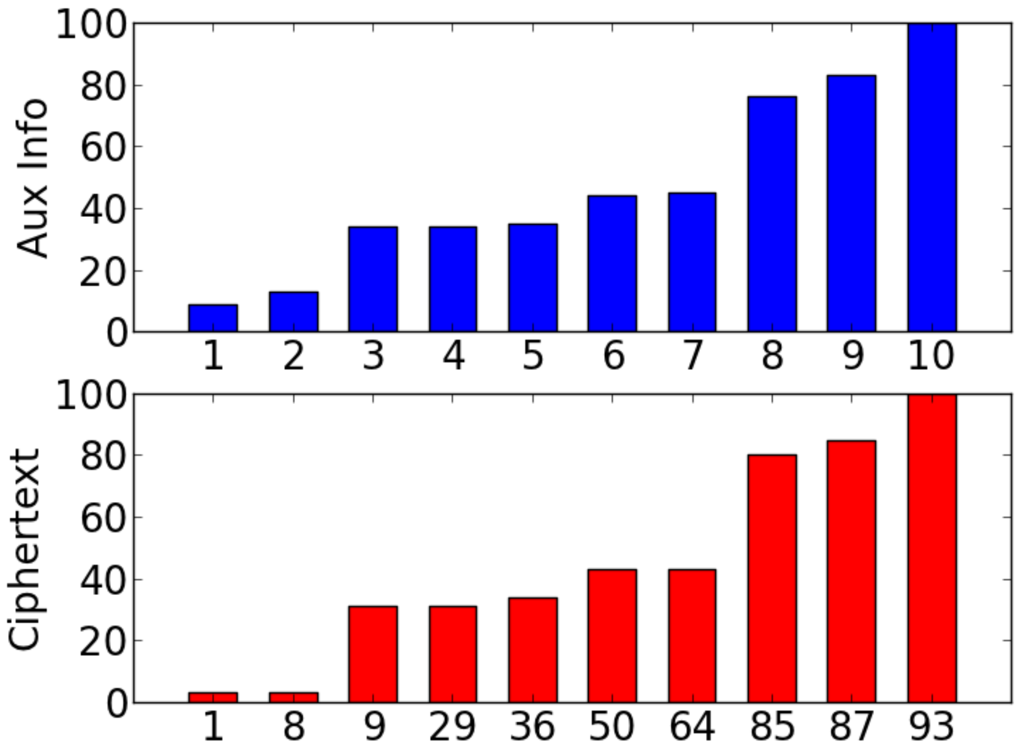
- Formulate the adversary's task as a **Linear Sum Assignment Problem (LSAP)**

- Use efficient solvers to find the answer
  - Hungarian algorithm – O(n^3)
  - Linear programming

Inference Attacks on OPE

- OPE reveals order of the plaintexts

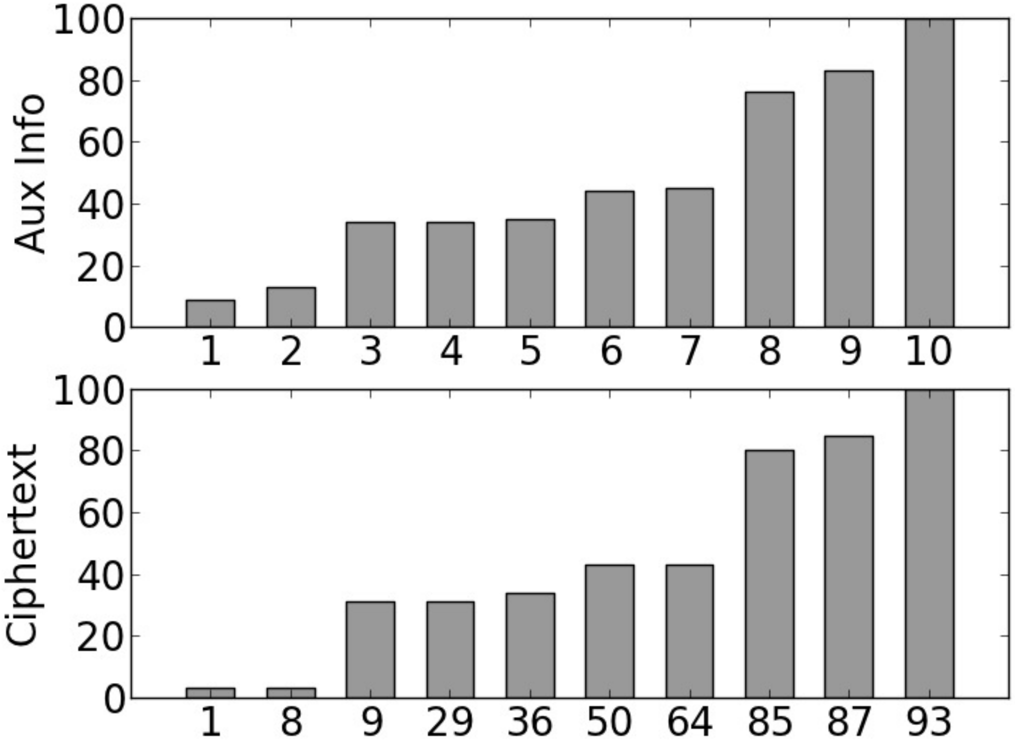
- Adversary can see the histogram AND the **cumulative frequencies**
  - ie, the cumulative sum of the histogram



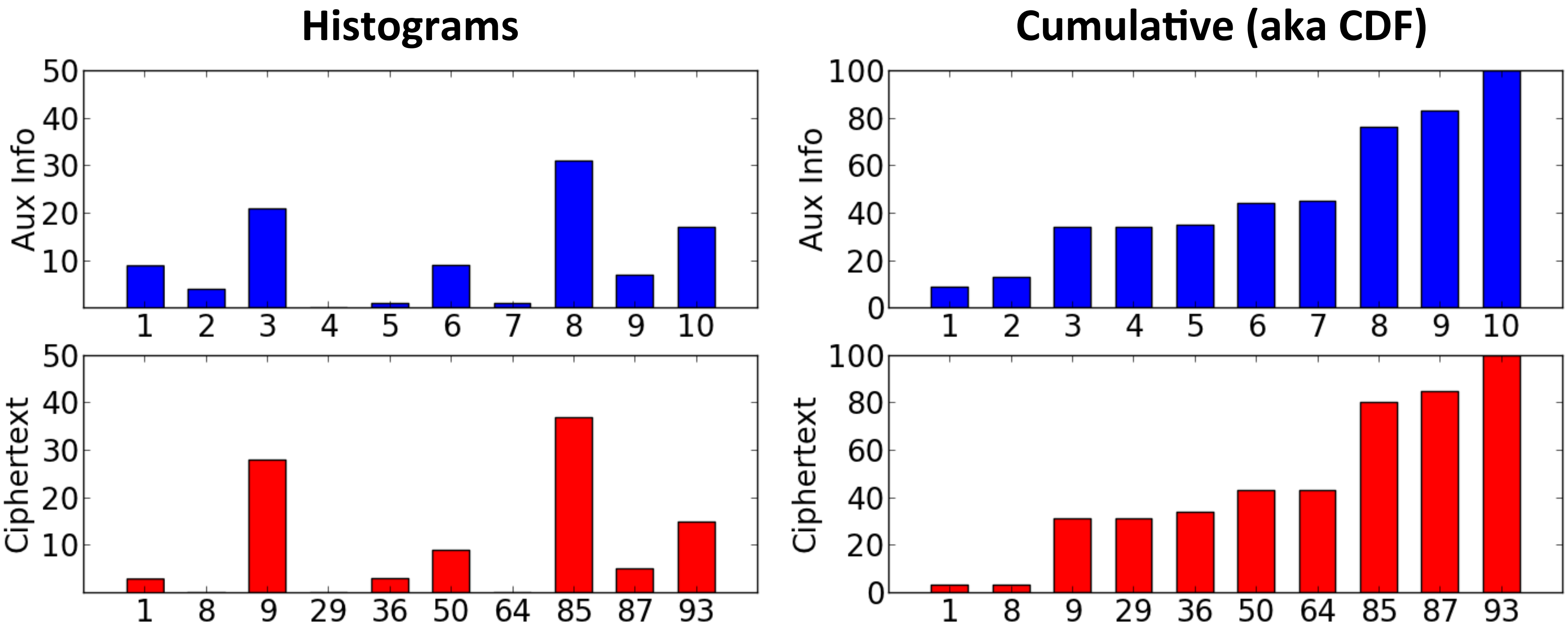
Sorting Attack

- Idea: If every value is present in the DB, then it's obvious which one is which

- Attack:
  1. Sort both sets into lexicographic order
  2. Match them up



Cumulative Attack



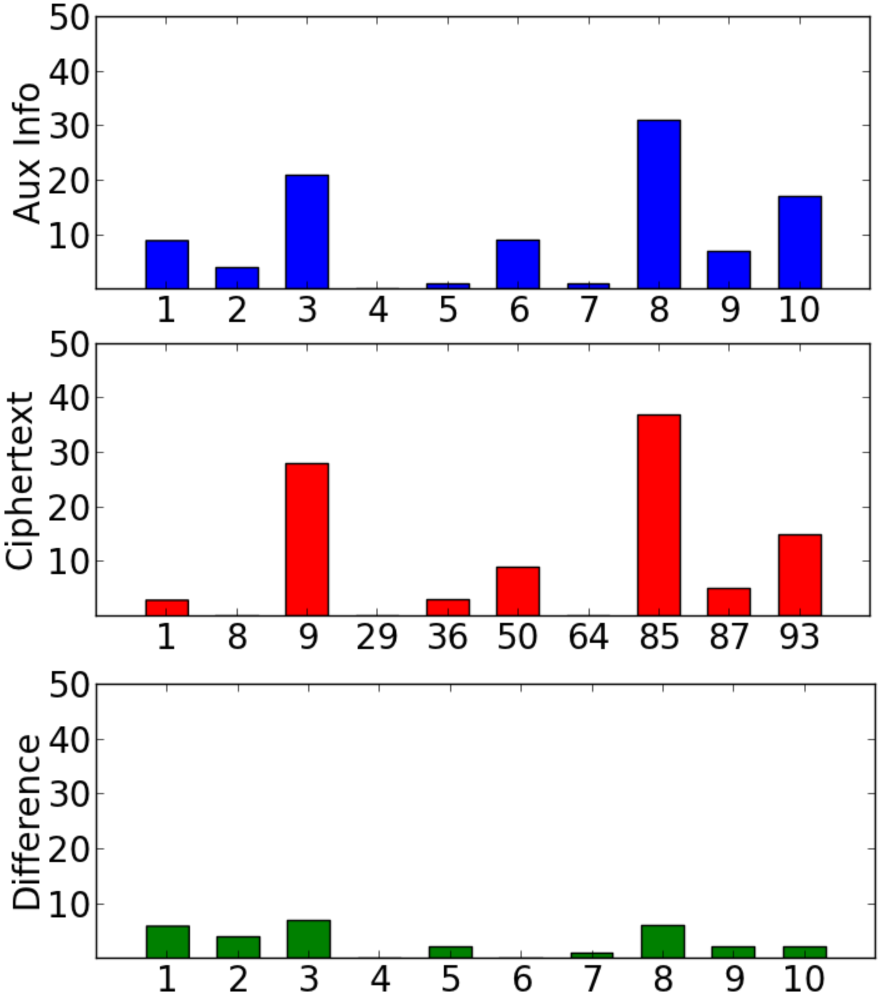
Histograms

Cumulative (aka CDF)

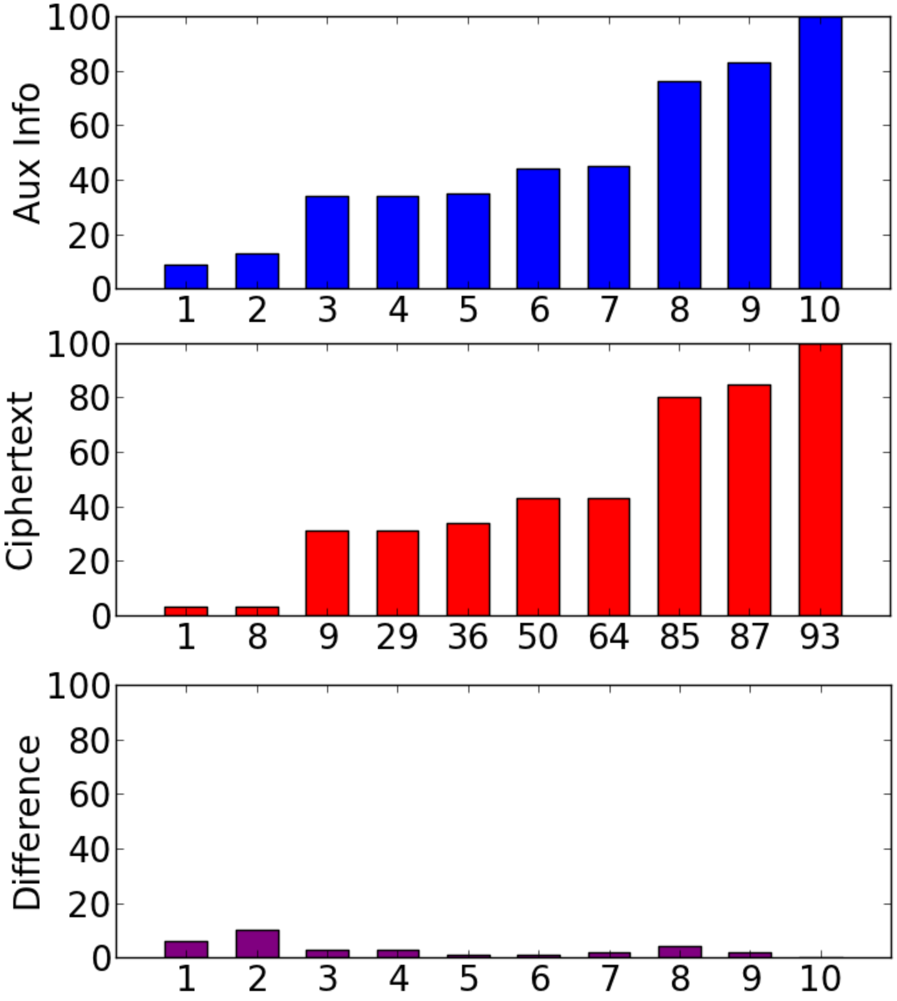
- Idea: Use both the histogram and cumulative frequencies to find the optimal matching

Cumulative Attack

Histograms



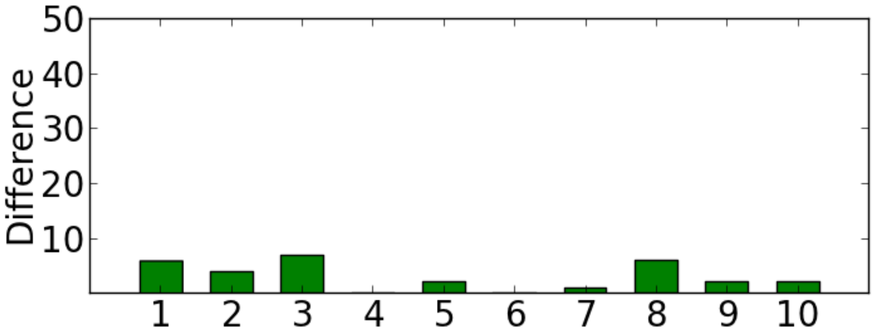
Cumulative (aka CDF)

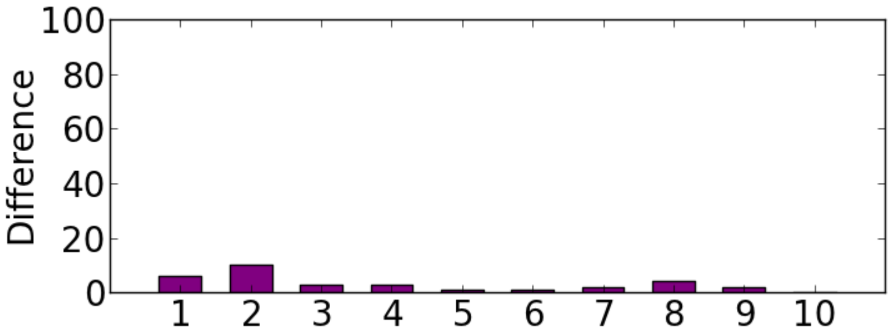


Cumulative Attack

- Include both vector differences in the LSAP

- Use the Hungarian algorithm to find the best solution that minimizes the differences





EMPIRICAL EVALUATION

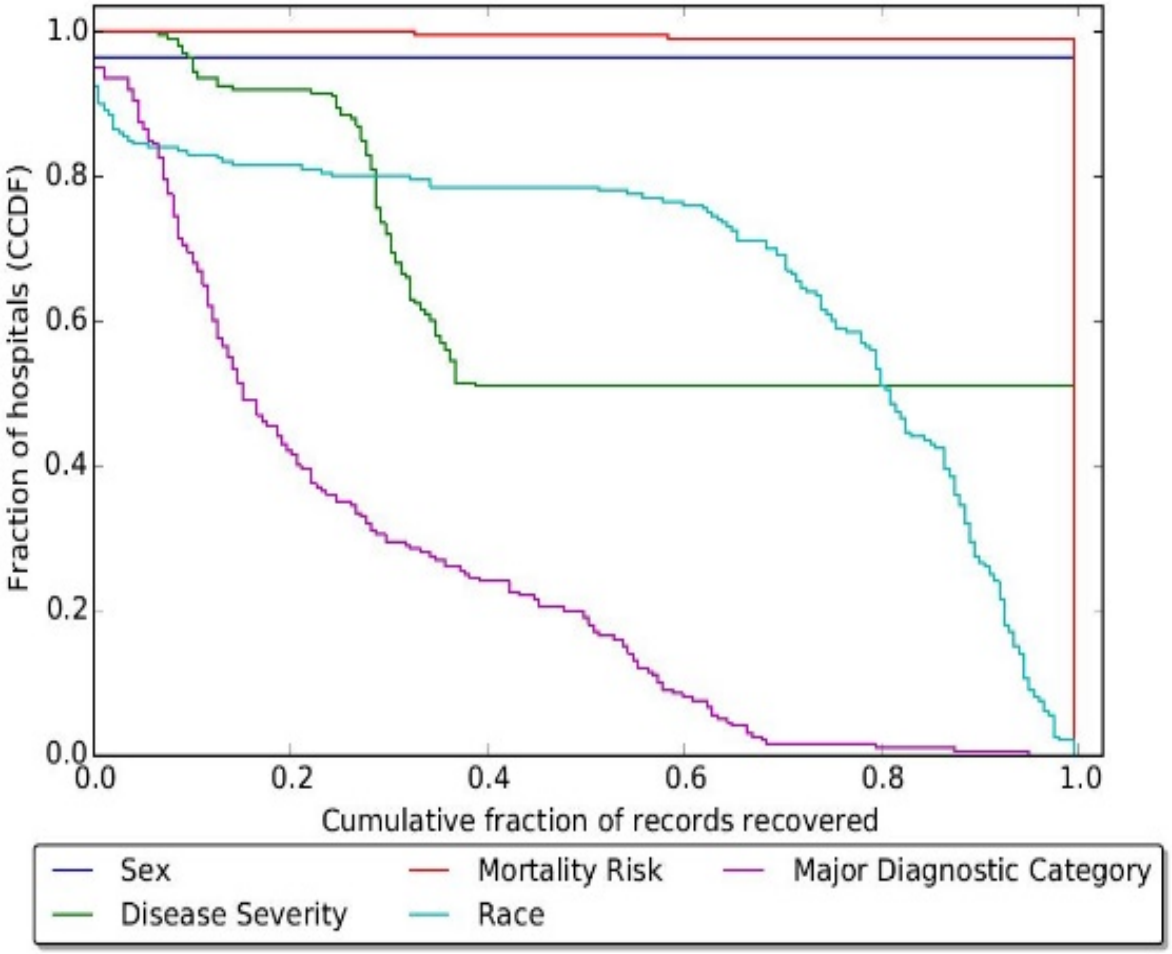
Experimental Setup

- Scenario: Medical data
- Application: electronic medical records (EMR)
- Target data: 2009 National Inpatient Sample (NIS) from Healthcare Cost and Utilization Project (HCUP)
- Auxiliary data
    - Texas Inpatient Public Use Data File (PUDF)
    - HCUP/NIS from 2004
- Attributes: sex, race, age, admission month, *patient died*, primary payer, length of stay, mortality risk, disease severity, *major diagnostic category*, admission type, admission source

L_p Optimization

- Mort. Risk: 100/99;
- MDC: 40/23;
- Dis. Sev.: 100/50;
- Race: 60/79:5

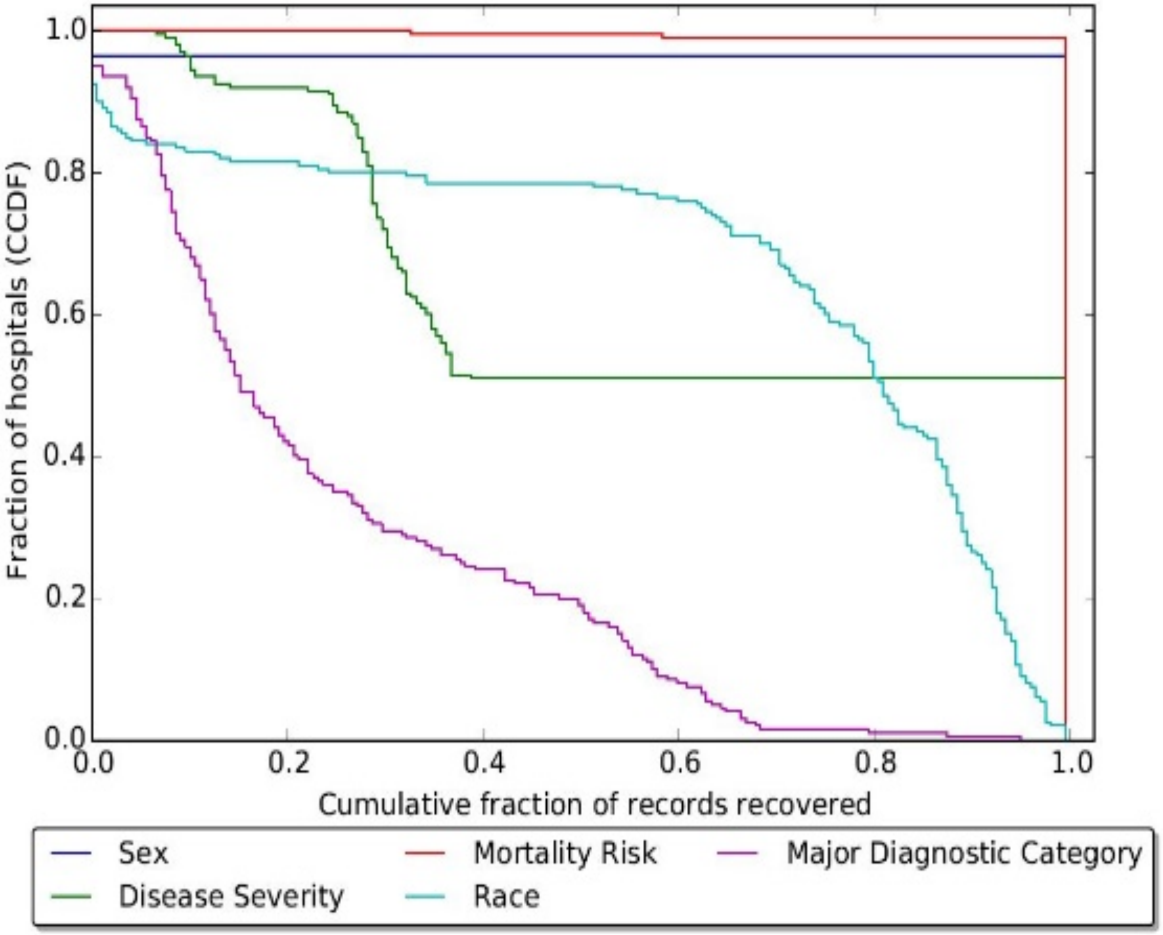
2004 HCUP/NIS vs. Texas PUDF



L_p Optimization

- Mort. Risk: 100/99;
- Patient Died: 100/100;
- MDC: 40/27:5;
- Dis. Sev.: 100/51;
- Race: 60/69:5

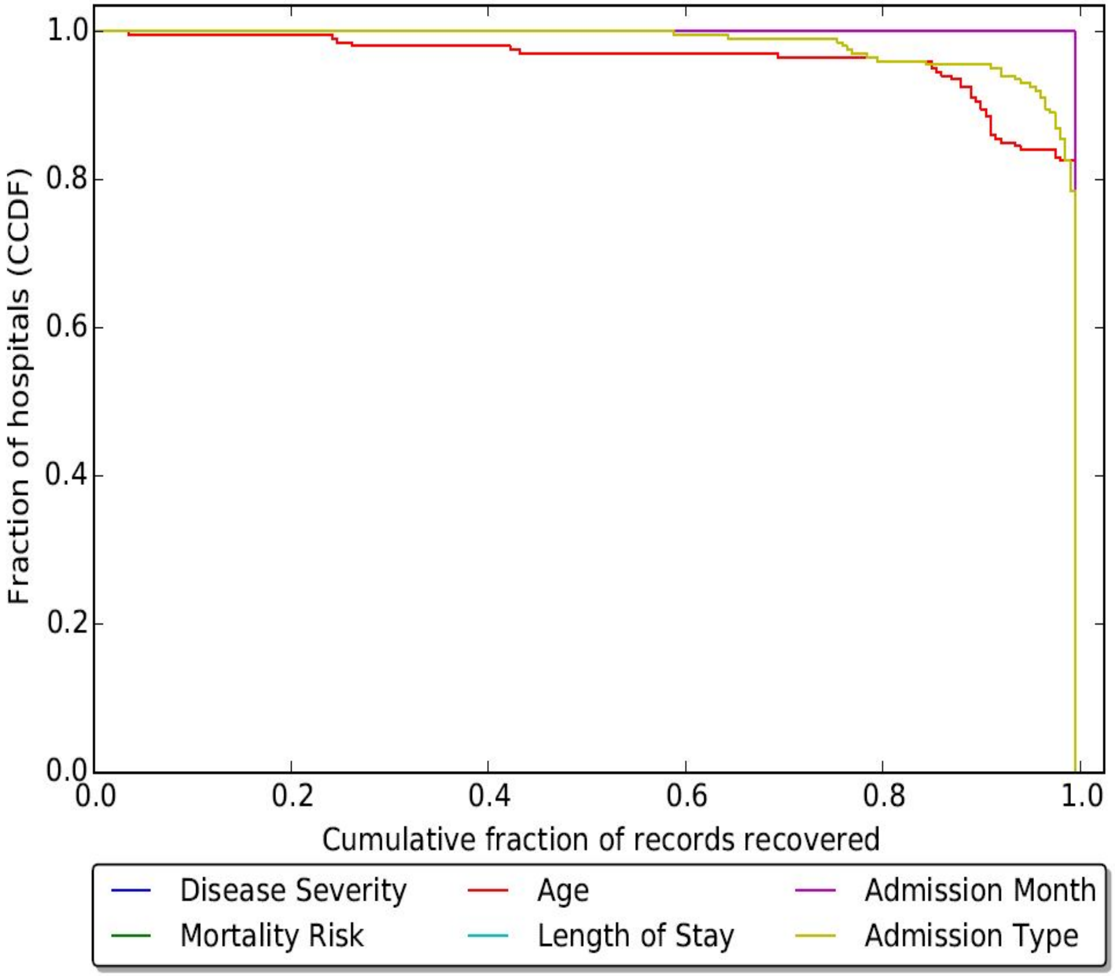
2009 vs. 2004 HCUP/NIS



Cumulative Attack

Large 2009 vs. 2004 HCUP/NIS

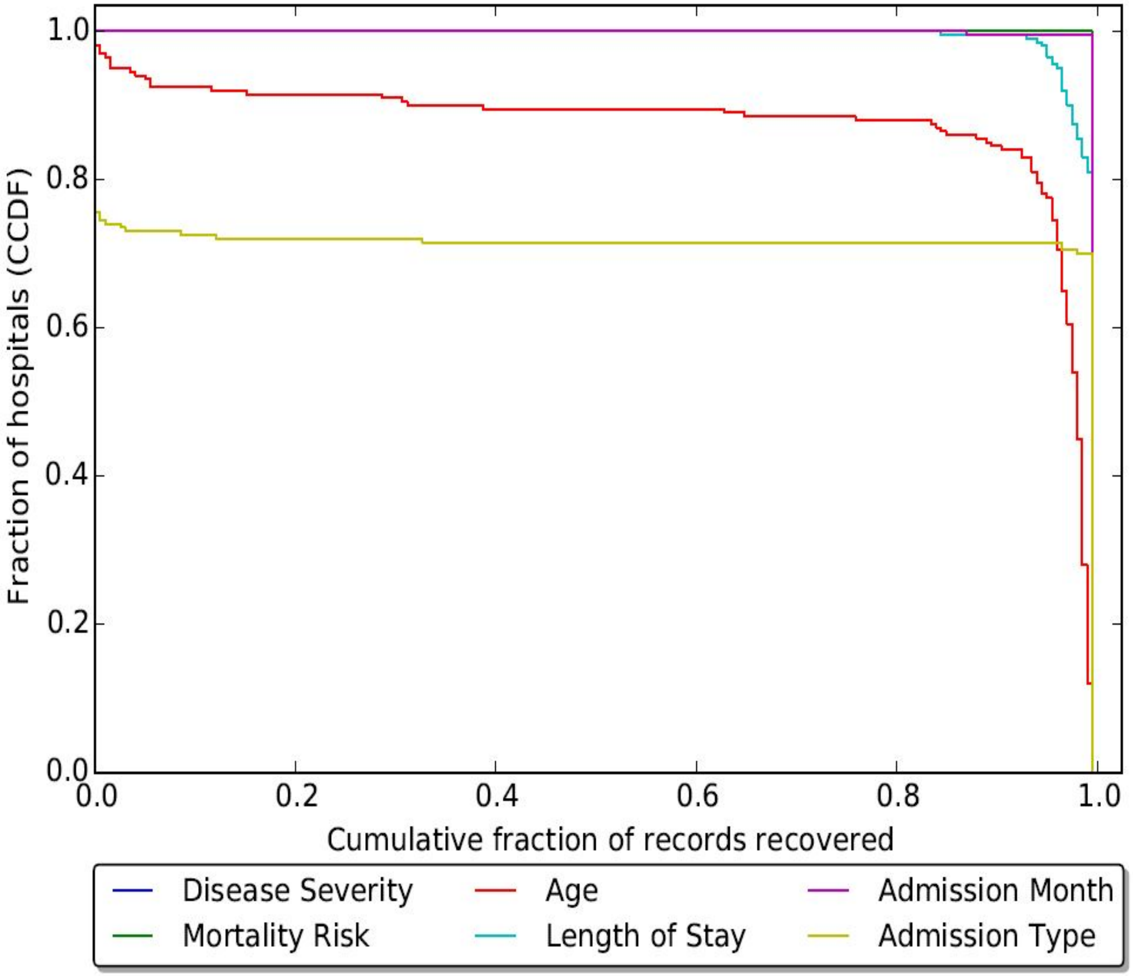
- Adm. Month: 100/100;
- Dis. Sev.: 100/100;
- Mort. Risk: 100/100
- LoS: 99.77/100;
- Age: 99/82:5;
- Adm. Type: 100/78:5



Cumulative Attack

Small 2009 vs. 2004 HCUP/NIS

- Adm. Month: 100/99:5
- Dis. Sev.: 100/100;
- Mort. Risk: 100/100
- LoS: 95/98;
- Age: 95/78;
- Adm. Type: 100/69:5



Reception

- Three of the projects cited were happy with our work
  - One publicly acknowledged and thanked us
  - Other asked to collaborate
  - Third used our work to motivate new research

- One project disputes our results

https://eprint.iacr.org/2015/979

DISCUSSION

Open Questions

- Lp Optimization vs Frequency Analysis?
  - Upcoming work with Moataz, Naveed, Kamara

- How well do these results generalize?
- What, if any, real data is safe for PPE?
  - New results coming soon!

- How can we build better systems?

How can we build better systems?

- Option 1 – Bite the bullet, live with the leakage
  - Ouch!
- Option 2 – Abandon PPE techniques altogether
  - Focus on other constructions, special hardware, etc…
- Option 3 – Develop (heuristic) defenses for PPE
  - Exciting! And fraught with peril!
  - Is this even feasible? Can PPE schemes be saved?
  - How do we measure success? How do we define security?
  - How do we assess the remaining risk?