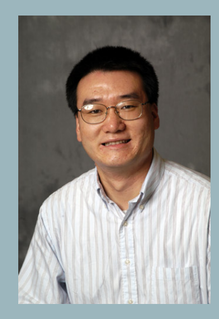
Bridging The Gap between Theory and Practice in Data Privacy

Ninghui Li (Purdue University)



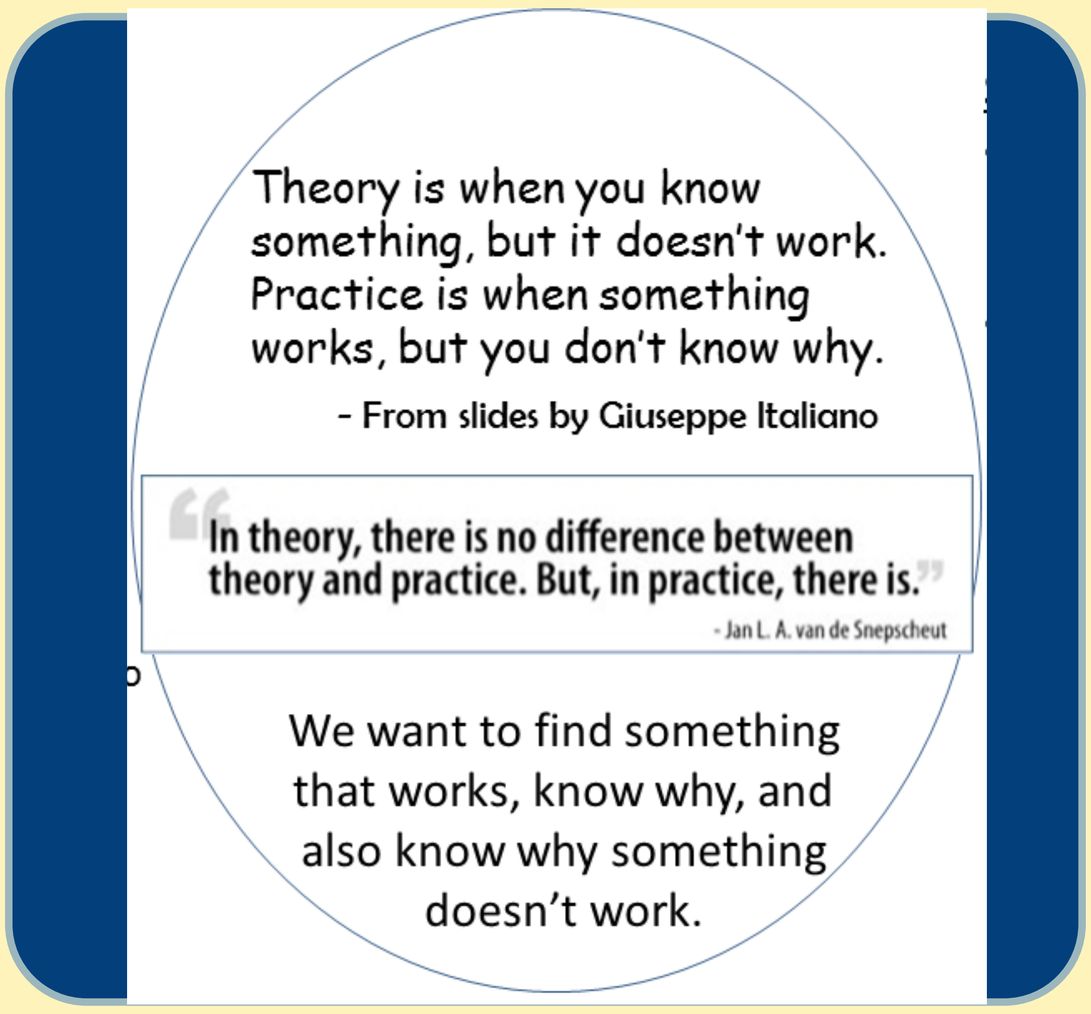
The objective of this project is to bridge the gap between theory and practice in data sharing and analysis while satisfying Differential Privacy. We aim to develop effective algorithms, and methodologies for combining theoretical analysis with experimental validations, focusing on concrete (instead of asymptotic) analysis where constants are spelled out.

The theoretical approach of proving asymptotic utility bounds is very limited

- Asymptotic analysis ignores constants (and oftentimes poly-logarithmic terms as well), which are critical for utility in practice.
- Only certain algorithms can be analyzed.
- Utility bound must hold for all datasets, including pathological ones, and are thus so loose as to be meaningless in practice.

A pure experimental approach is also limited

- Accuracy of an algorithm often depends on many parameters, such as nature of data sets, algorithmic parameters.
- Easy to reach conflicting conclusions when comparing algorithms by choosing different parameters.



Approach

- Develop a concrete approach to understanding the utility
- Identify different sources of errors for an algorithm, come up with approximate estimations of their impact, and empirically verify analysis of errors
- Divide an algorithm into a combination of multiple ideas, and empirically analyze the effect of different ideas.
- Build an arsenal of techniques for effective DP algorithms.

Local DP setting

- Each individual perturbs data before submitting; a server recovers frequencies of values. Used by Google and Apple.
- We developed a framework for such protocols that yields a simple recovery algorithm, and a formula of estimation of variance.
- Our optimized protocols reduce variances over state-of-art methods by orders of magnitude.

Classification with high-dimensional data

- Existing DP algorithms for classification does not scale to data with hundreds or more dimensions.
- Many algorithms (such as those based on stochastic gradient descent, random matrix projection, genetic programming) perform poorly empirically.
- A simple algorithm based doing private greedy local search seems to perform the best.

Other tasks with high-dimensional data

- Frequent itemset mining
- Answering marginal queries

Future research topics

- Develop practically relevant lower-bound results
- Better understanding of relaxation of DP

Interested in meeting the PIs? Attach post-it note below!

National Science Foundation
WHERE DISCOVERIES BEGIN

The 3rd NSF Secure and Trustworthy Cyberspace Principal Investigator Meeting
January 9-11, 2017
Arlington, Virginia

PURDUE UNIVERSITY