Modeling, prediction and diagnosis for network security

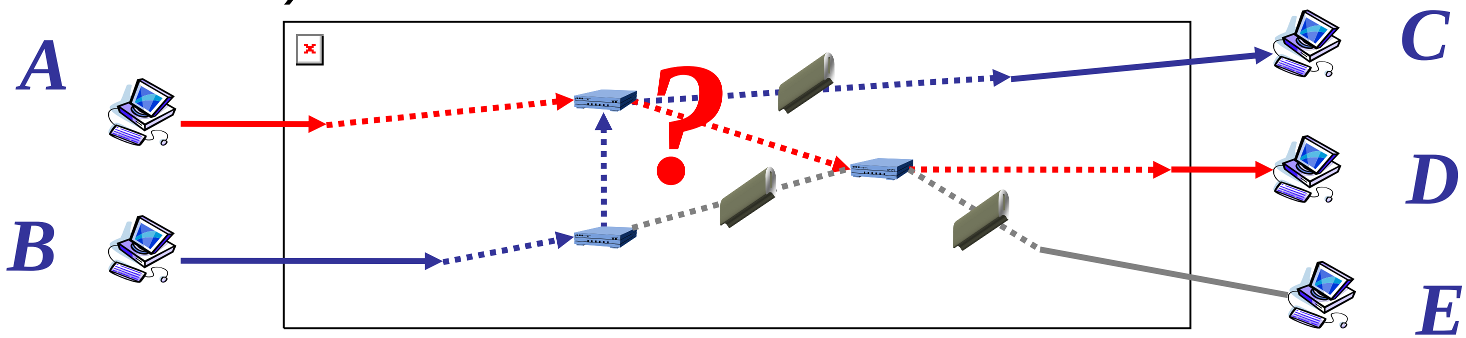
Alfred Hero

University of Michigan

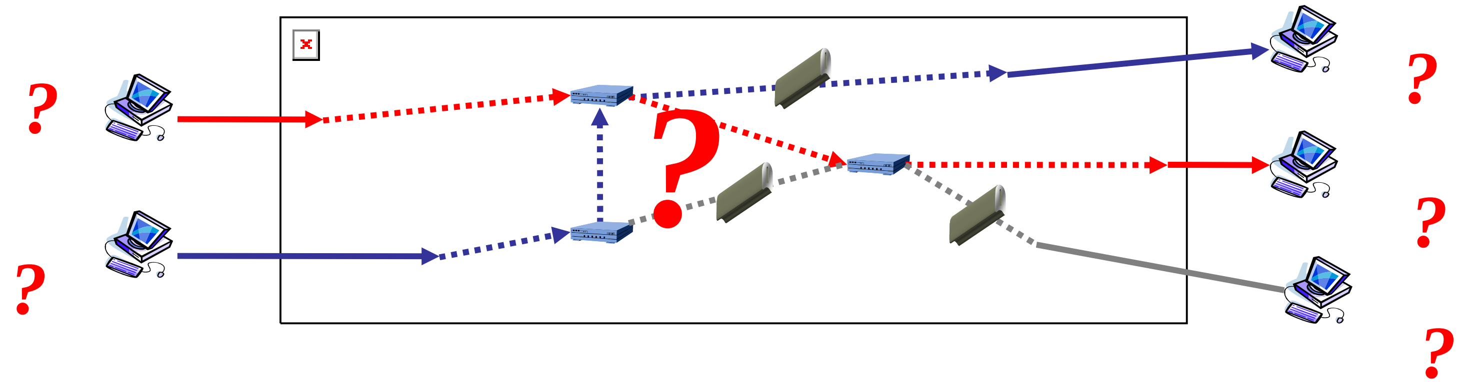
1. Network monitoring and tomography
2. Science of security: opportunities
3. Concluding remarks

1. Network monitoring and tomography

- Internally sensed network tomography (Treichler05, Rabbat06)



- End-point prediction and tracking(Justice06)



2. Science of security: opportunities

- Scientific method
  - Observation
  - Hypothesis
  - Prediction
  - Experiment
  - Evaluation

- Science of Security
  - Sparse, incomplete?
  - Model selection?
  - Baseline drift?
  - Observer effect?
  - Benchmarks?

Observation

- Challenge: Critical security breaches are covert, rare, and non-repeatable
  - Any set of observations will necessarily be sparse and incomplete
  - Persistent and pervasive multimodal monitoring impractical

Cross-fertilizations

- Information-driven sensor management
  - Plan-ahead learning with POMDP (Carin:06, Blatt06)
  - Q-learning for reactive targets (Kreucher:06)
  - Performance prediction (H07, Castanon08)
- ISNT applications (Rabbat08,Justice06), but more research needed
  - Necessary and sufficient sampling rate?
  - Distributed processing and inference?
  - Scalable algorithms and approximations?

Hypothesis

- Challenge: infer stable models of attack and ambient behaviors that can be reliably tested

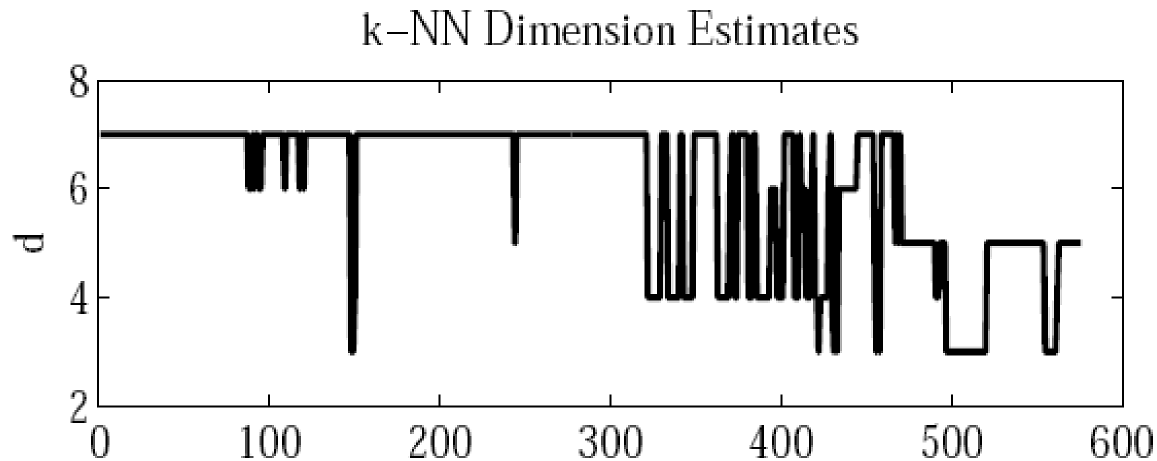
  – Central question: how to discover hidden latent structure of partially observed variables?

Cross-fertilizations

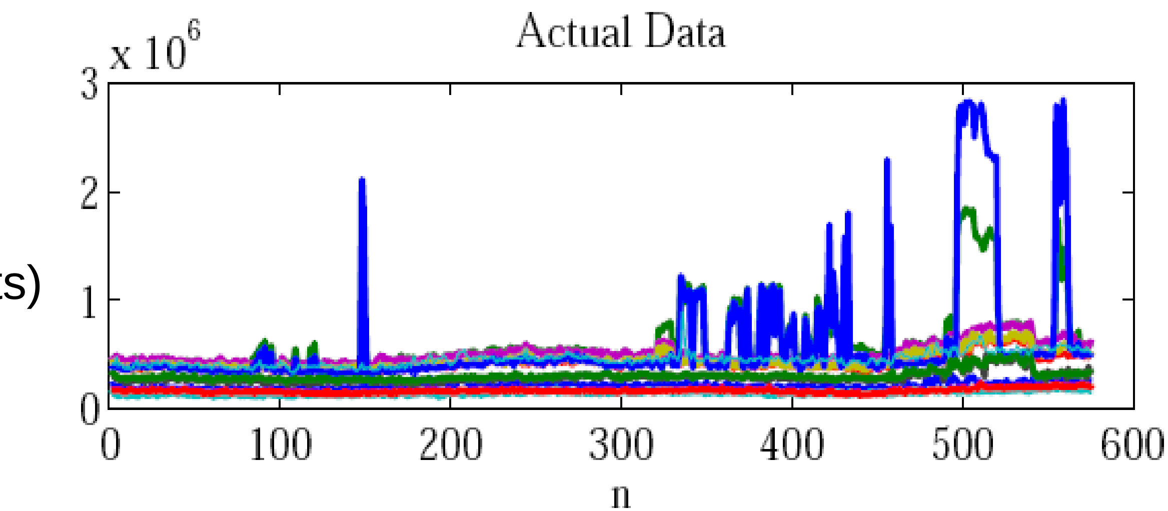
- Statistical model selection: how many attack patterns are there and how to identify them?
- Unsupervised hypothesis generation
  - Bayesian factor analysis (West05)
  - Information driven PCA (FINE, IPCA) (Carter08_b)
  - Complexity filtering (Carter08_a)
  - Social networks of behavoir (Xu09)
- How to make these approaches scalable to whole network security applications?

Complexity filtering (Carter:08_a)

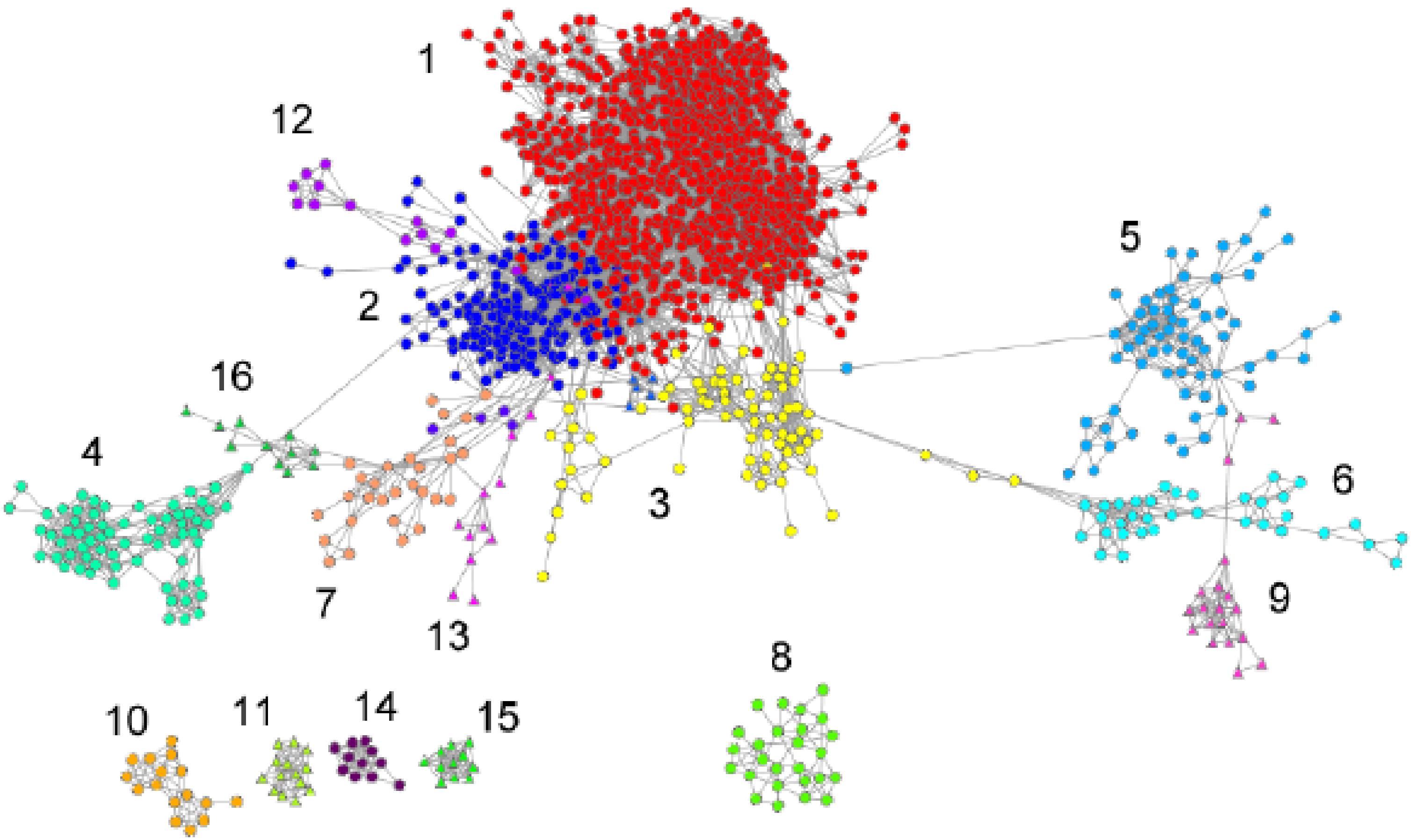
Intrinsic dimension estimator



Abilene Netflow data (Total number packets)

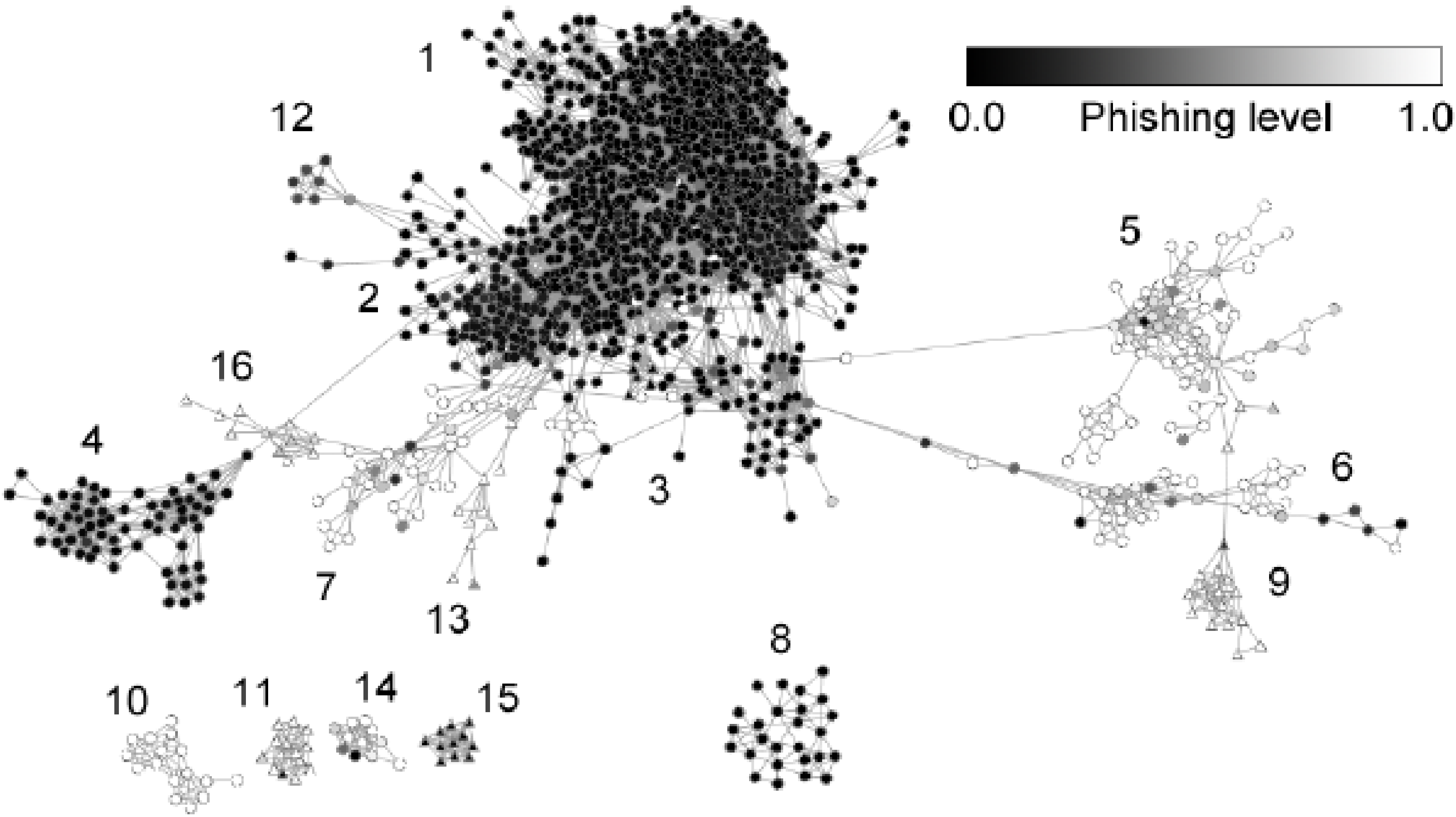


SocNet of SPAM harvestors (Xu:09)



Results from October 2006 using similarity in spam server usage
(visualization created using Cytoscape)

SocNet of SPAM harvestors (Xu:09)



Results from October 2006 colored by phishing level

Prediction

- Challenge: learn truly predictive and generalizable models that
  - Track dynamic shifts over time or space
  - Extract information from high dimension
  - Integrate uncalibrated diverse data types

Cross-fertilizations

- Predictive anomaly detection
  - Transductive learning (Scott08)
  - Geometric entropy minimization (H06)
- Flexible graphical/topological models
- How to make these methods scalable?
  - decomposable version of Lakhina04's PCA for whole-network diagnosis

Dynamic dwMDS for Abilene (Patwari:05)

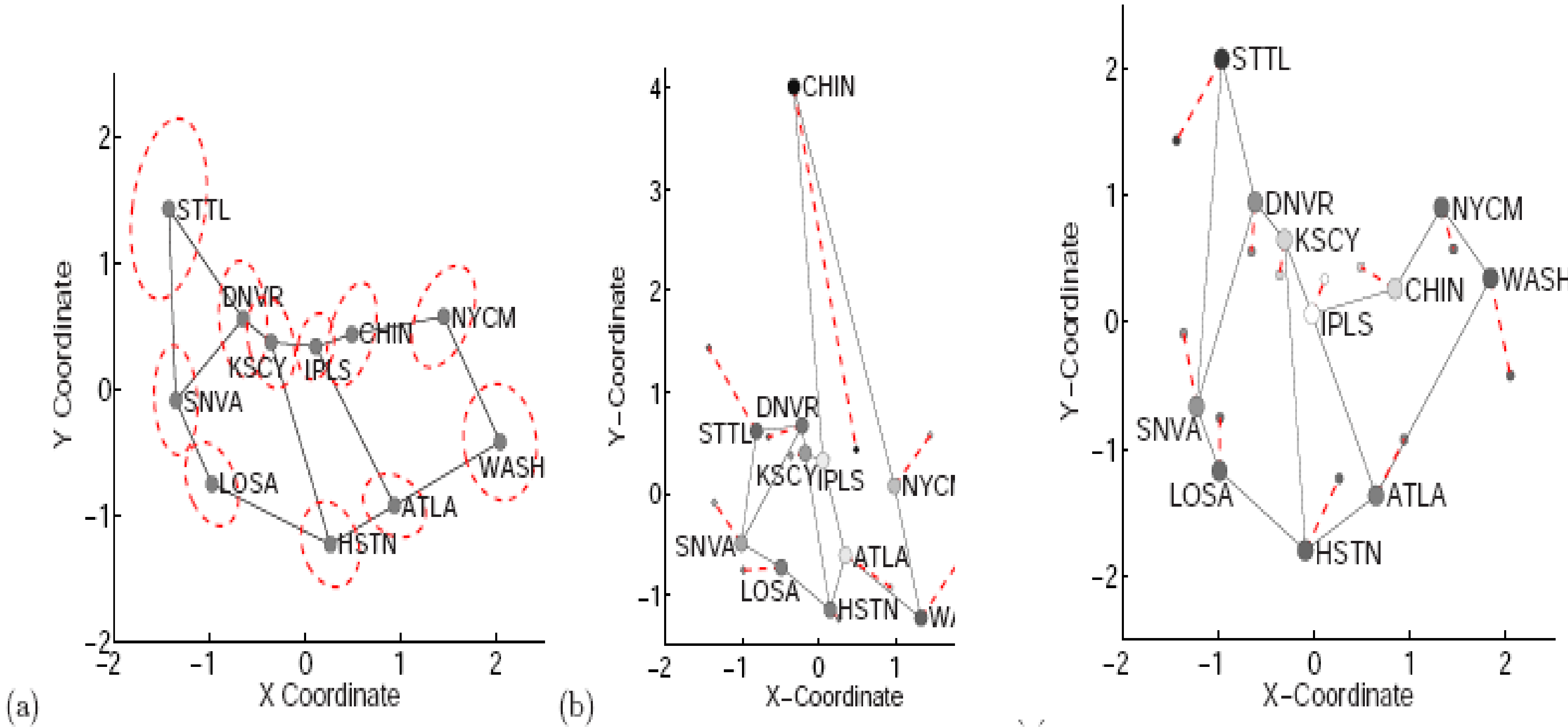


Figure 2: (a) Mean (•) and 1-σ uncertainty ellipse (- - -) of router maps from 2-Jan to 29-Jan. Maps during (b) port scan on 6-Jan 17:55 and (c) attack on 20-Jan 01:00, show router coordinates (•) connected (- - -) to the mean (·) from (a), and shaded by error value e_i. All figures show Abilene backbone links (—).

Experiment

- Challenge: simulation relies on stale or speculative models while real-world data collection is difficult due to
  - Disruption of infrastructure
  - Unreliable ground truth
  - Significant “observer effects”

Cross-fertilizations

- Adversarial experiment design approaches
  - Dynamic generalizations of adversarial classification (ACRE, Lowd&Meek06, Dalvi04) and greedy minimax (Kraus07)
  - RL w observer effect (Kreucher06, Murphy06)
- Design of experiments for medical clinical trials have similar constraints

Evaluation

- Challenge: establish reliable methods of on-line and offline performance prediction
  - Incomplete label information/ground truth
  - Curse of dimensionality
    - require order $1/e^p$ samples to determine the values of p experimental variables within error e

Cross-fertilizations

- Bayesian meta-analysis: what is posterior uncertainty of predicted estimation error?
- DOE benchmarking: what is theoretically attainable algorithm performance?
  - Coding and information theory
    - Error exponents, Fano, Rate-Distortion bounds
    - Tradeoffs between security and usability (H03)
  - Minimax, maximax and minimin performance prediction: function estimation and imaging (BickelRitov:90,KostolevTsybakov:93)

3. Final remarks

- Developing a Science of Security is challenging.
- Leverage from other disciplines with high throughput data
  - Image reconstruction and tomography
  - Social networks and economic behavior models
  - Genetics, immunology, and epidemioogy
- Main open problems
  - Adversarial learning environment
  - Rapidly changing baseline
  - Data impoverishement
  - Scalable plan ahead sampling