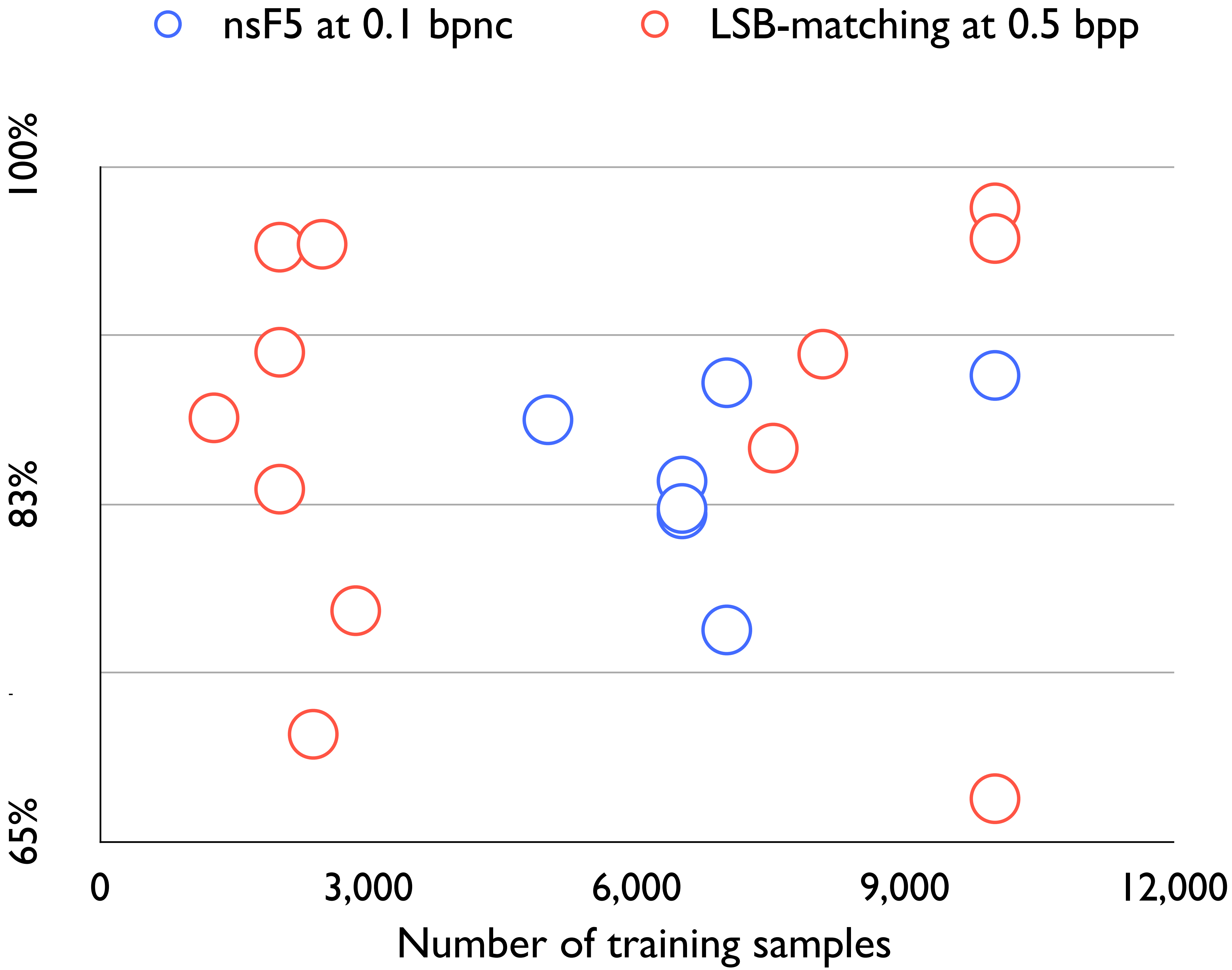
UNIVERSITY OF OXFORD

Ivans Lubenko and Andrew Ker

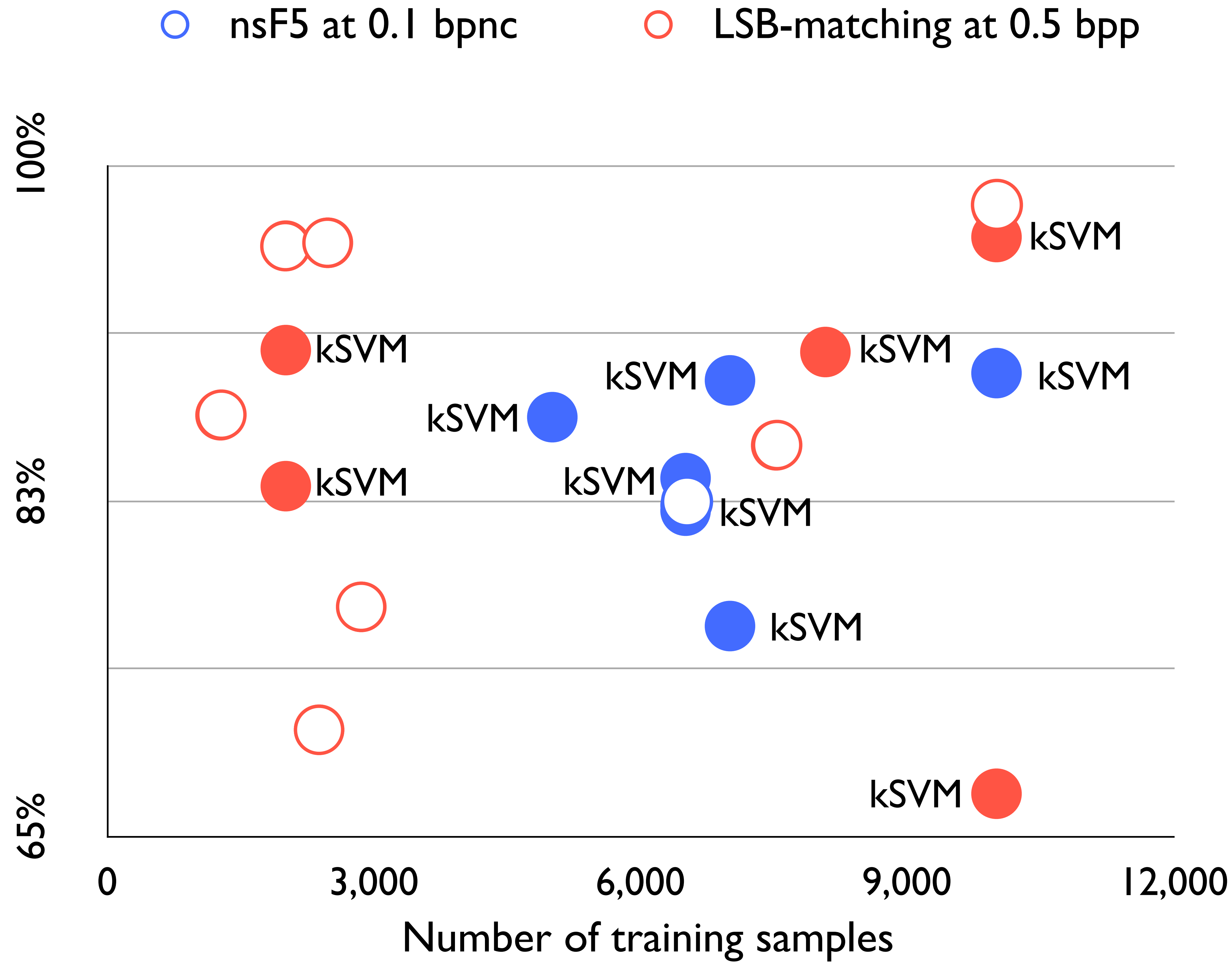
Going from Small to Large Data in Steganalysis

25 January 2012 @ IS&T/SPIE Electronic Imaging

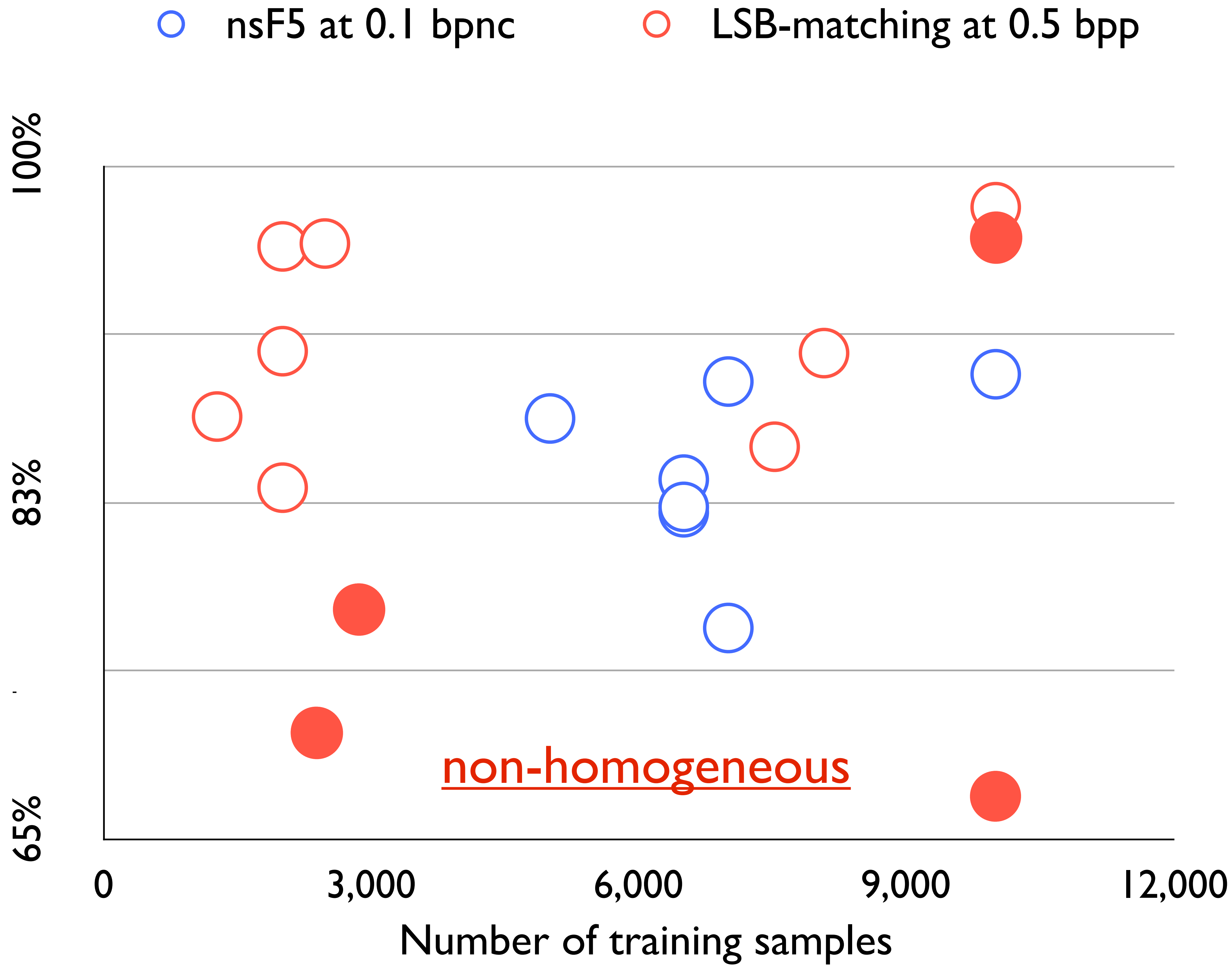
Going from Small to Large Data in Steganalysis



Going from Small to Large Data in Steganalysis



Going from Small to Large Data in Steganalysis



non-homogeneous

OBSERVATION

- Complex classifiers demonstrate low accuracy on *non*-homogeneous data

EXPLANATION

- They overfit the source (not training set)

HYPOTHESIS

- *Simple classifier trained on large data will work better*

- DATA SET:

  - 800,000 JPEGs x 2 (cover vs. nsF5)
  - 4000 different uploaders
  - collected from public sources
  - is non-homogeneous and difficult:

  86.9% using kSVM with CC-PEV on subset of 5,000 examples with cover/nsF5 at 0.1 bpnc

Going from Small to Large Data in Steganalysis

- FEATURES:

  - CC-C300 from [1]
  - large dimensionality:
    48,600 features
  - likelihood of linear separability
  - slow to train (no kSVM in [1])
  - large to store:

48,600 x 2 x 800,000 x 8 bytes = 620GB

[1] Jan Kodovsky, “Steganalysis in high dimensions: fusing classifiers built on random subspaces”, 2011

ONLINE ALGORITHMS

a hot topic in Machine Learning for last few years

- process one training example at a time
- one pass through data

+ unlimited training
+ no parameter tuning
+ low memory requirement

AVERAGE PERCEPTRON

Decision rule:

dot product

predicted label

$$y(x) = sign(w_{avg}^T x)$$

Simple update rule:

true label of x

$$w_i = w_{i-1} + x_i t_i$$

training example

- very fast
- regularised via averaging:

average weight vector

$$w_{avg} = w_{avg} + w_i$$

OUR EXPERIMENTS

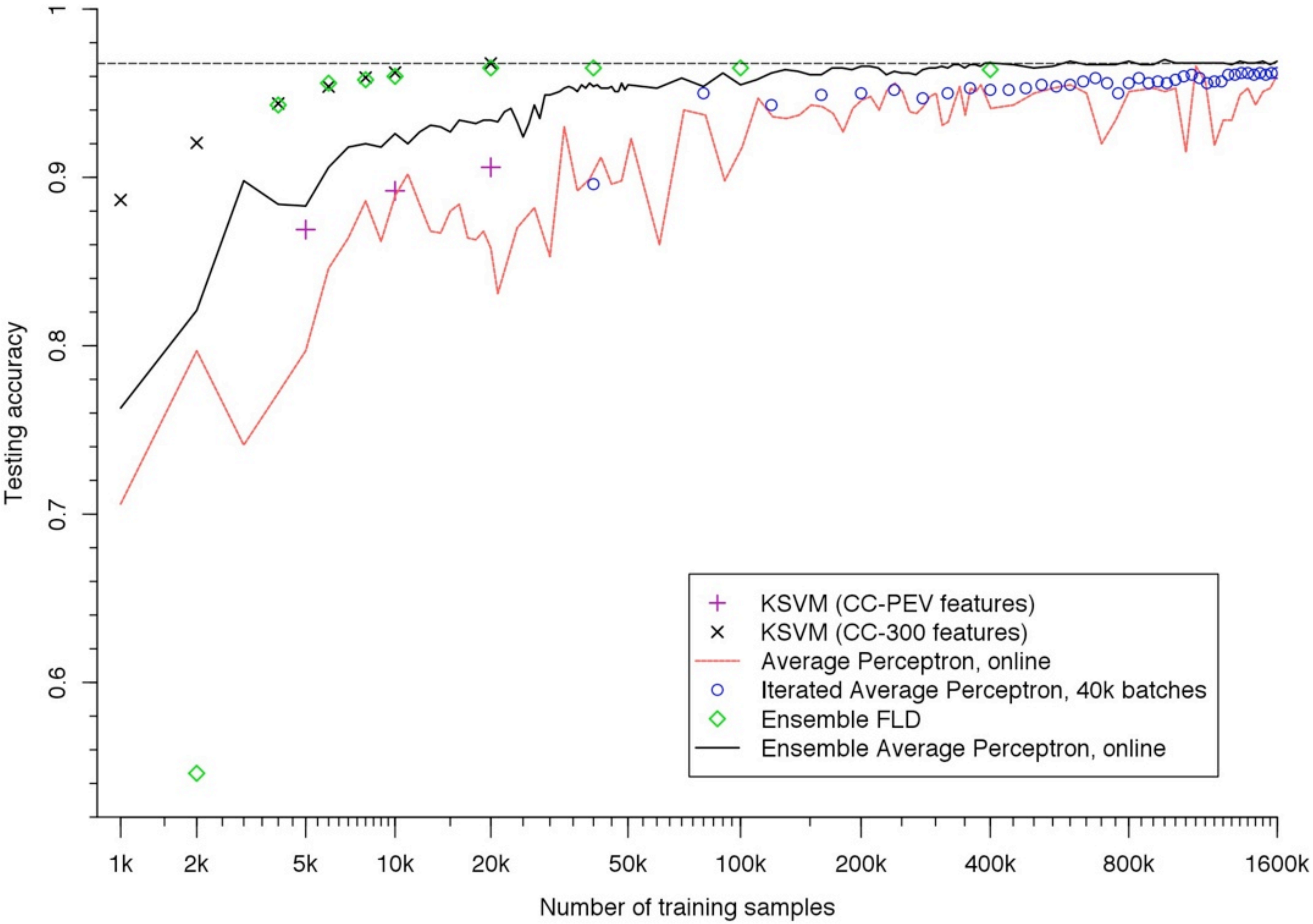
cover vs. (no-shrinkage) nsF5

0.1 and 0.2 bpnc

1.6 million training examples

separate testing set of 20,000 examples

Going from Small to Large Data in Steganalysis



OUR EXPERIMENTS

1. KSVM - over 10 days on 20,000
2. Ensemble FLD - over 7 days on 400,000
3. Online Average Perceptron - 1 hour
4. Iterated Average Perceptron - 2.5 hours
5. Ensemble Online Average Perceptron - 7h

BIG DATA is:

1. large *training set* + large *feature set*
2. more *important* than complex classifier

   linear algorithms as accurate as complex algorithms on small data

3. *very fast*

   using online algorithms

4. required for *non-homogenous* data classification

Going from Small to Large Data in Steganalysis

- FUTURE DIRECTIONS:
  - More stable simple online algorithms
  - Non-linear online algorithms
  - How large data works with small features
  - Active learning in steganalysis