On the Feasibility of Author Identification in the Era of Big Data

Neil Zhenqiang Gong
EECS, UC Berkeley
Arvind Narayanan, Hristo Paskov (CS, Stanford), Neil Gong, John Bethencourt, Emil Stefanov, Eui Chul Richard Shin and Dawn Song (EECS, UC Berkeley)

Motivation

- Anonymous/pseudonymous contents are everywhere!

Motivation

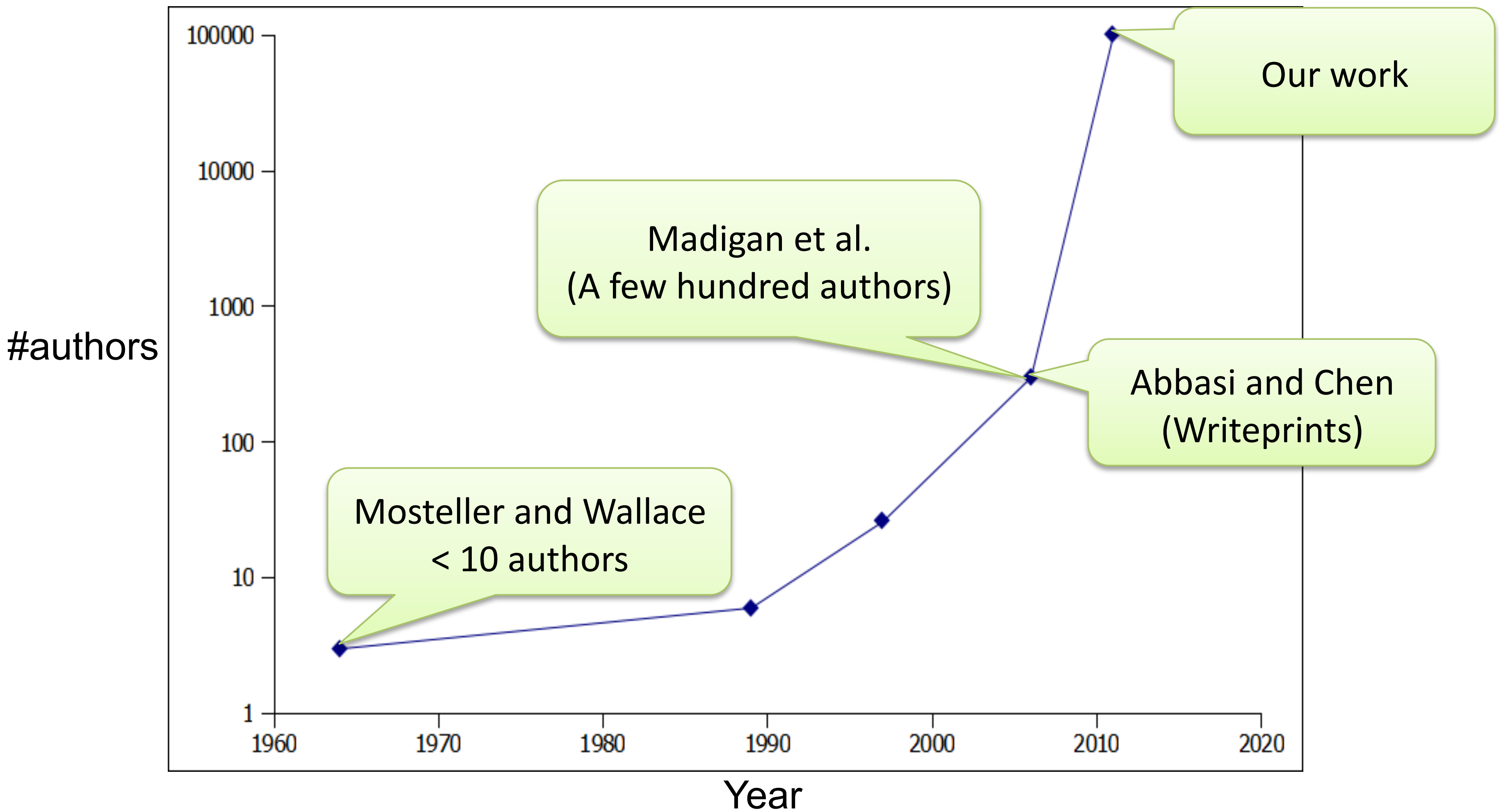
- Anonymous contents:
  - Sensitive political topics
  - Sensitive personal psychological/health issues.
- Identifying authors = huge privacy attack!
- Possible via writing style at Large scale?

Notable Coups for Stylometric Author Identification

Shakespeare-Bacon controversy in 19th century

Disputed Federalist Papers ~50 years ago

Graph of #authors vs. year



Author identification behaves qualitatively different at large scale.

Threat Model

Attacker: oppressive government, etc.

Authors are *not* protecting themselves

Use author ID as first step

Follow up with other methods:

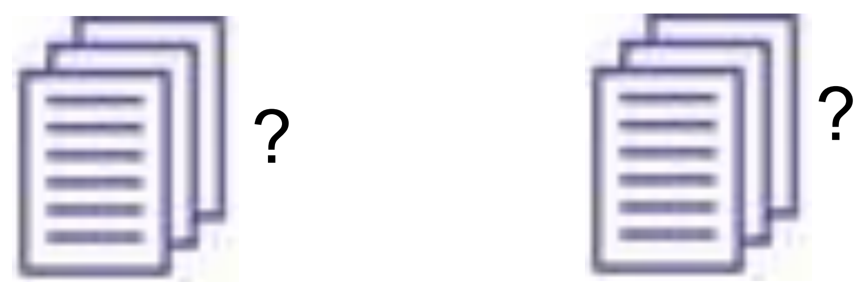
topic, viewpoints, location…

Problem Definition

- Given:
  - N authors

  - A set of labeled documents for each author.

- Target:
  - Identify the author of anonymous documents.



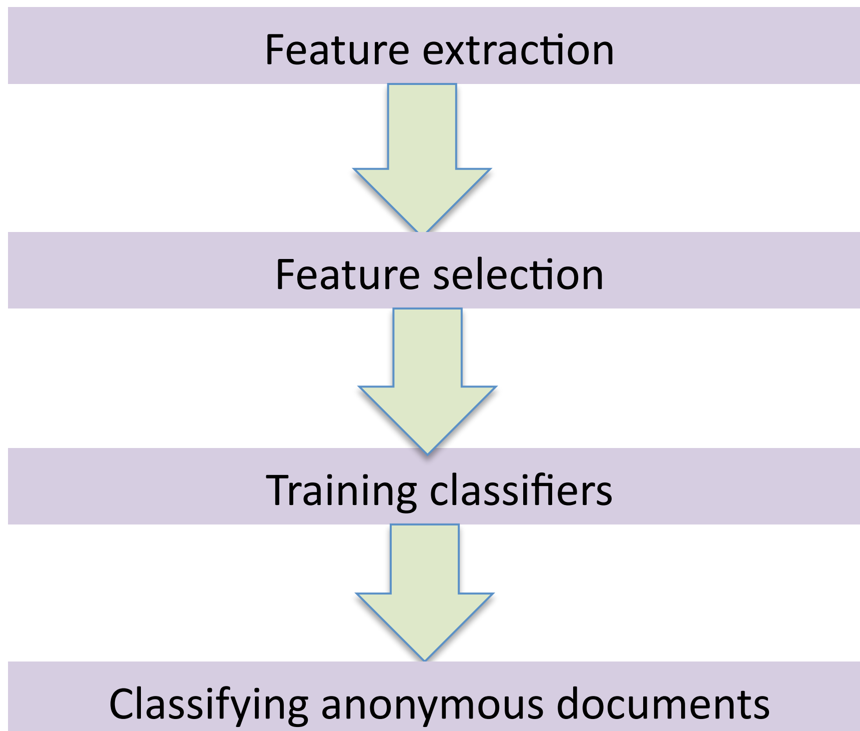
Approach

- Identification is a multi-class classification problem.
  - Classes: authors
  - Training examples: labeled documents
  - Test examples: anonymous documents

Roadmap

- Issues of large scale
- Dataset
- Experimental results
- Conclusion
- Future work

Machine Learning Framework



Scale impacts every part

Feature Extraction

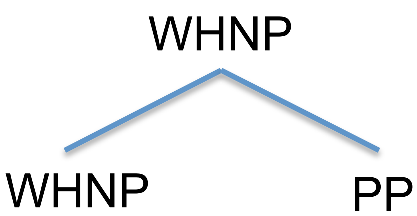
Writeprints features

Category	Description	Count
Length	number of words/characters in post	2
Vocabulary richness	Yule's K^2 and frequency of *hapax legomena*, *dis legomena*, etc.	11
Word shape	frequency of words with all upper-case letters, all lower-case, etc.	5
Word length	frequency of words that have 1–20 characters	20
Letters	frequency of *a* to *z*, ignoring case	26
Digits	frequency of 0 to 9	10
Punctuation	frequency of . ? ! , ; : ( ) " - '	11
Special characters	frequency of other special characters ` ~ @ # $ % ^ & * _ + = [ ] { } \ | / < >	21
Function words	frequency of words like 'the', 'of', and 'then'	293
Syntactic category pairs	frequency of every pair (A, B), where A is the parent of B in the parse tree	789

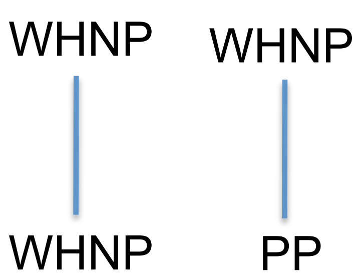
Our new feature

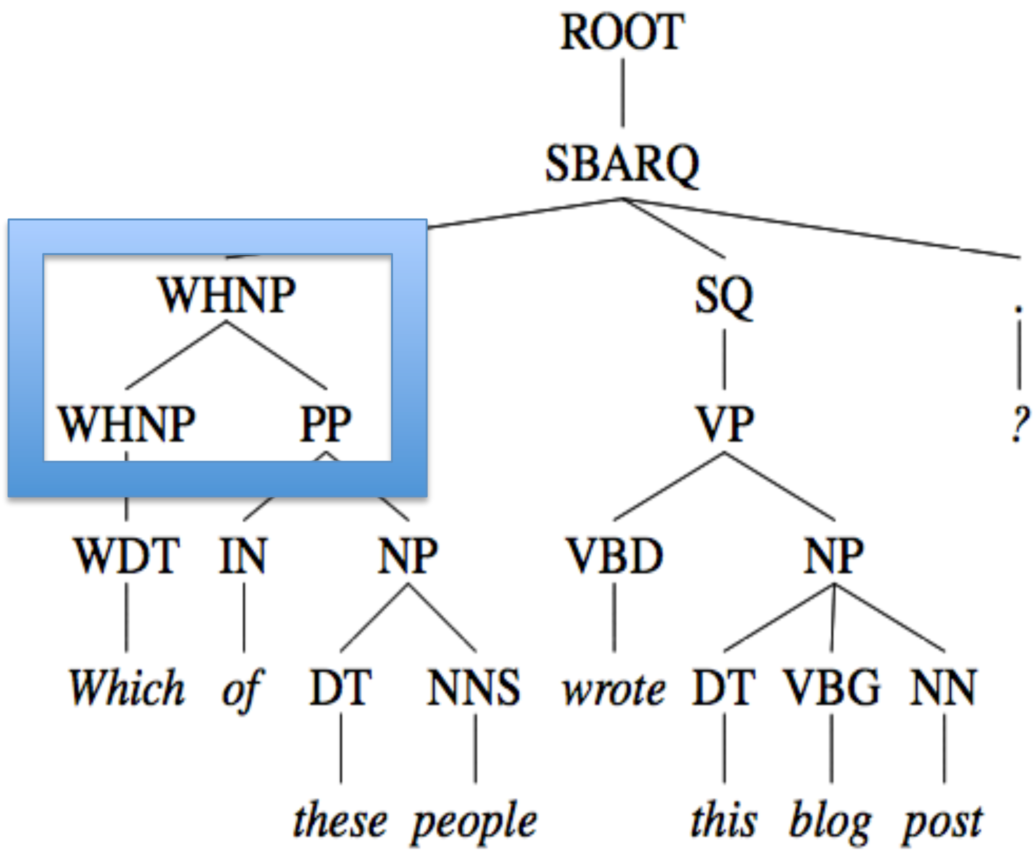
Syntactic Features

Previous feature:



Our feature:

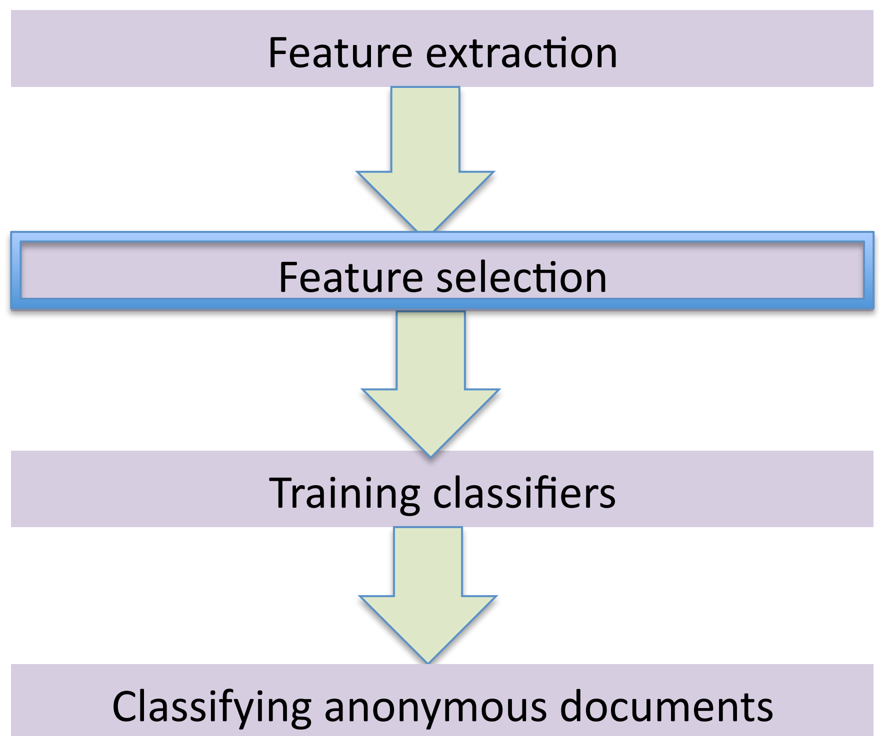




A sample parse tree produced by the Stanford Parser.

~1200 features!

Machine Learning Framework

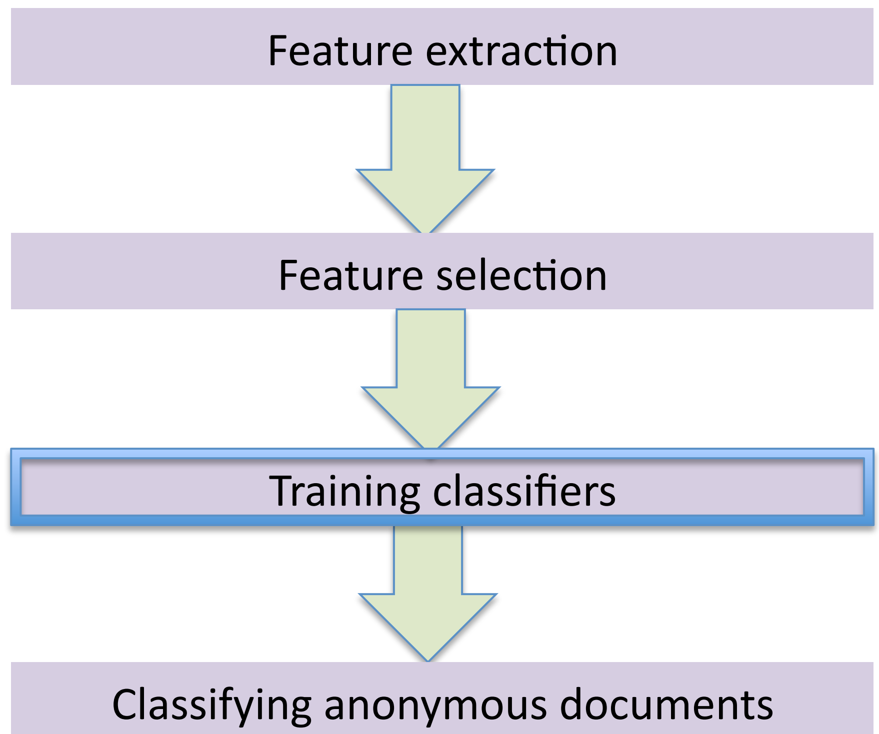


Feature Selection

- Information gain
- Document frequency

- Helpful for small scale
- Not helpful for large scale

Machine Learning Framework



Classifiers

- Nearest neighbor (NN)
- Naïve Bayes (NB)
- Support vector machines (SVM)
- Regularized least square classifier (RLSC)
- Ensemble classifier
  - NN + RLSC

Regularized Least Square Classifier (RLSC)

- Comparable accuracy to SVM
- Much more scalable than SVM
- One-vs-all
  - Training binary classifier for each author
- Class imbalance
  - Subsampling a small number of negative examples
  - Cost sensitive learning. ✔
    - Penalizing more for misclassifying positive examples

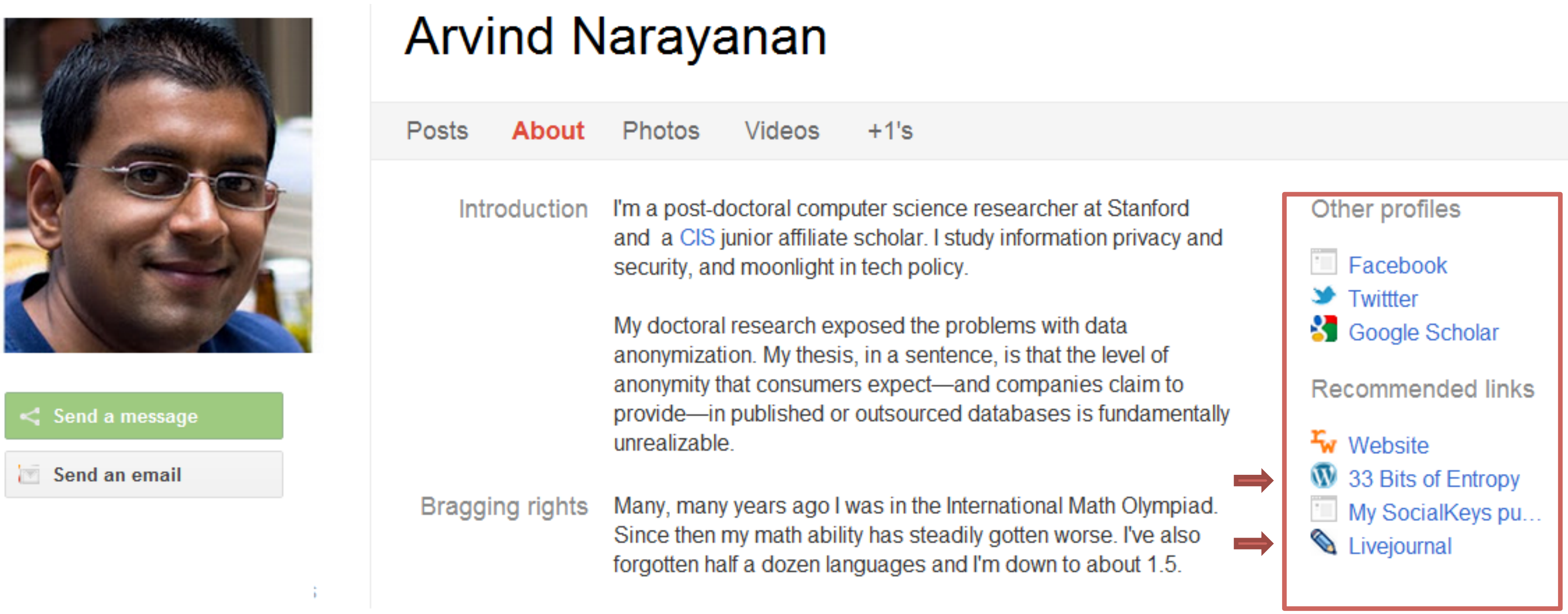
Dataset

spinn3r

ICWSM 2009 Dataset: ~94k blogs

Minimum 7,500 characters per blog
(roughly 8 paragraphs)

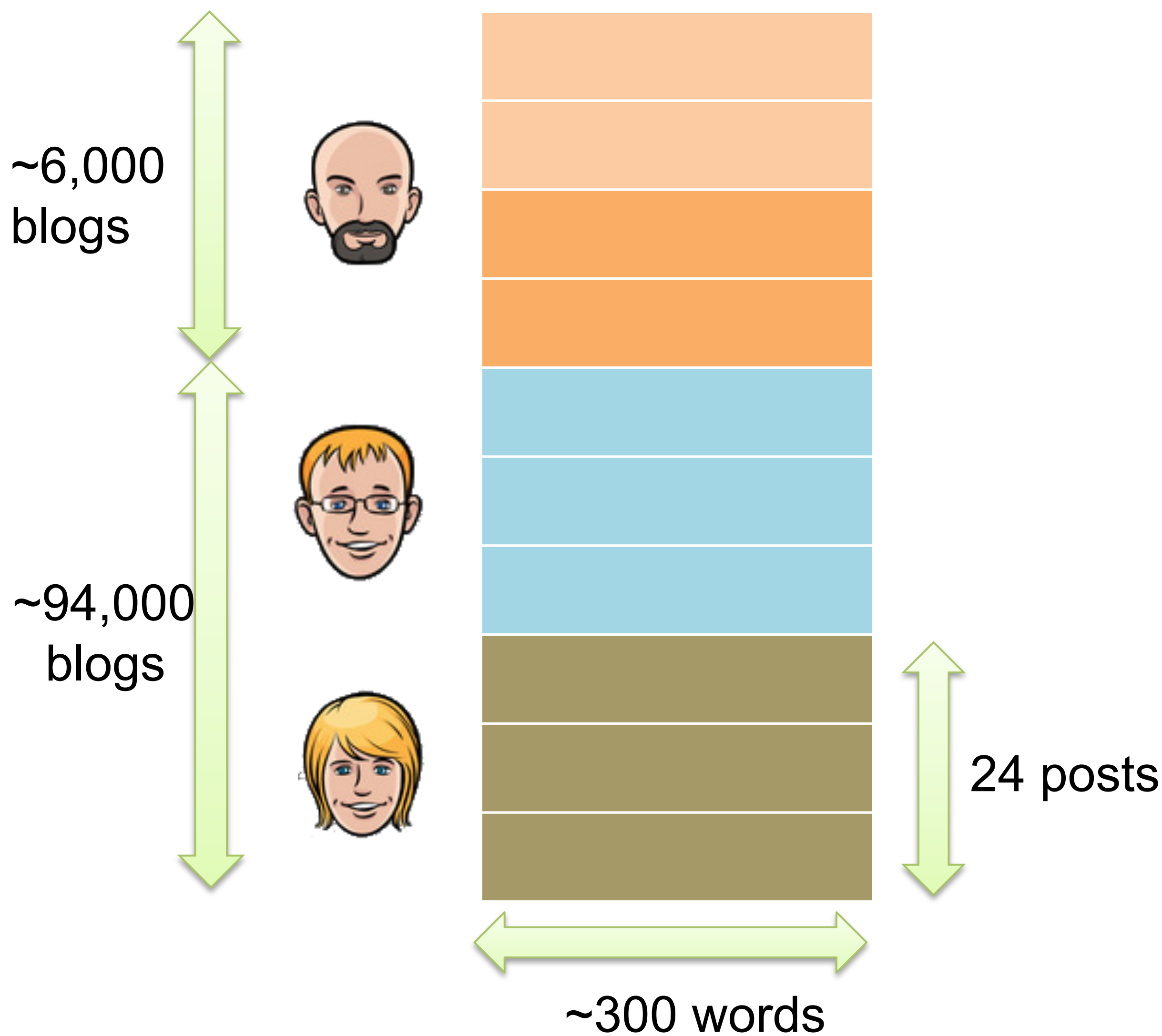
Dataset : Google Profiles



~6,000 blogs

~3,600 authors

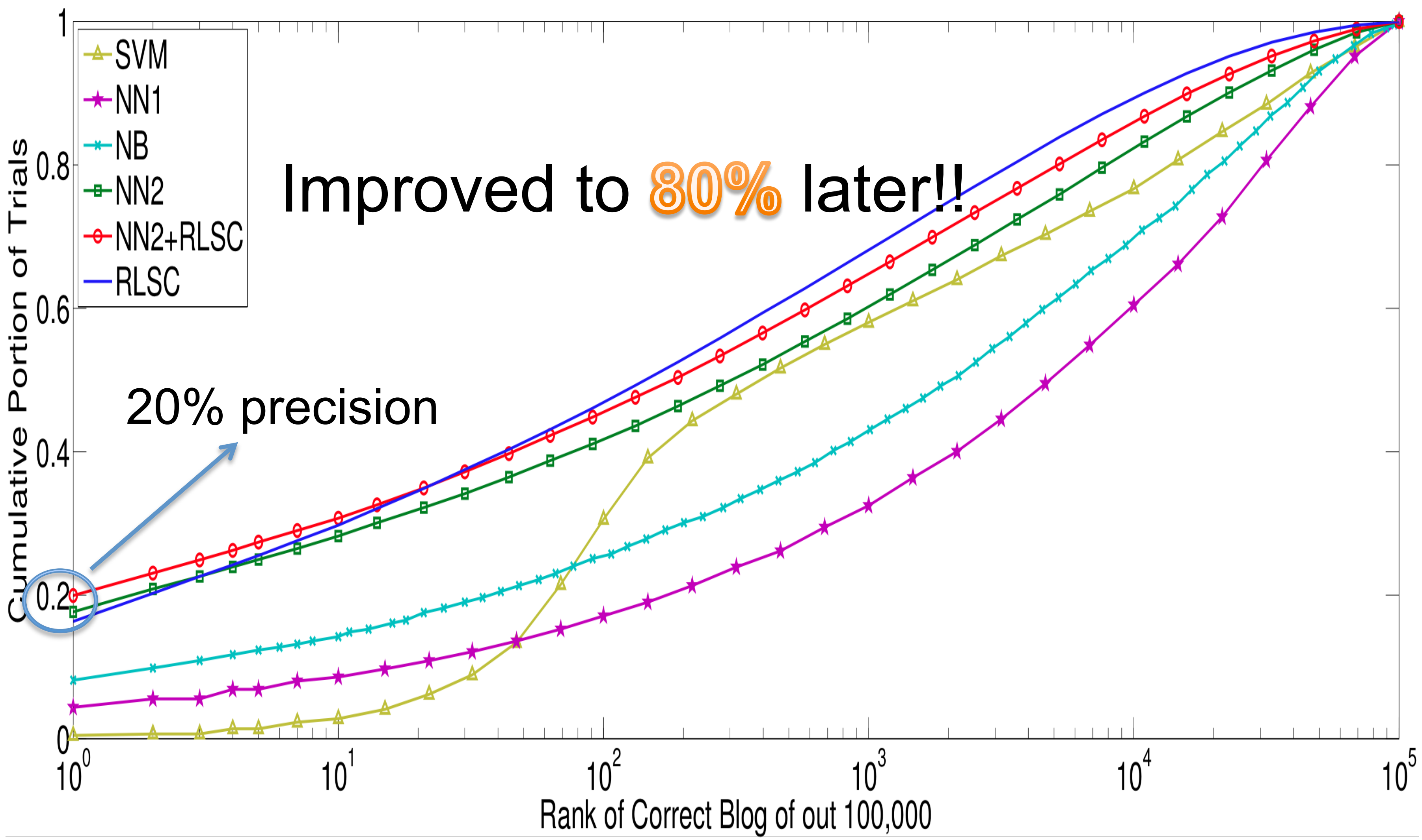
Data Size



Experimental Design

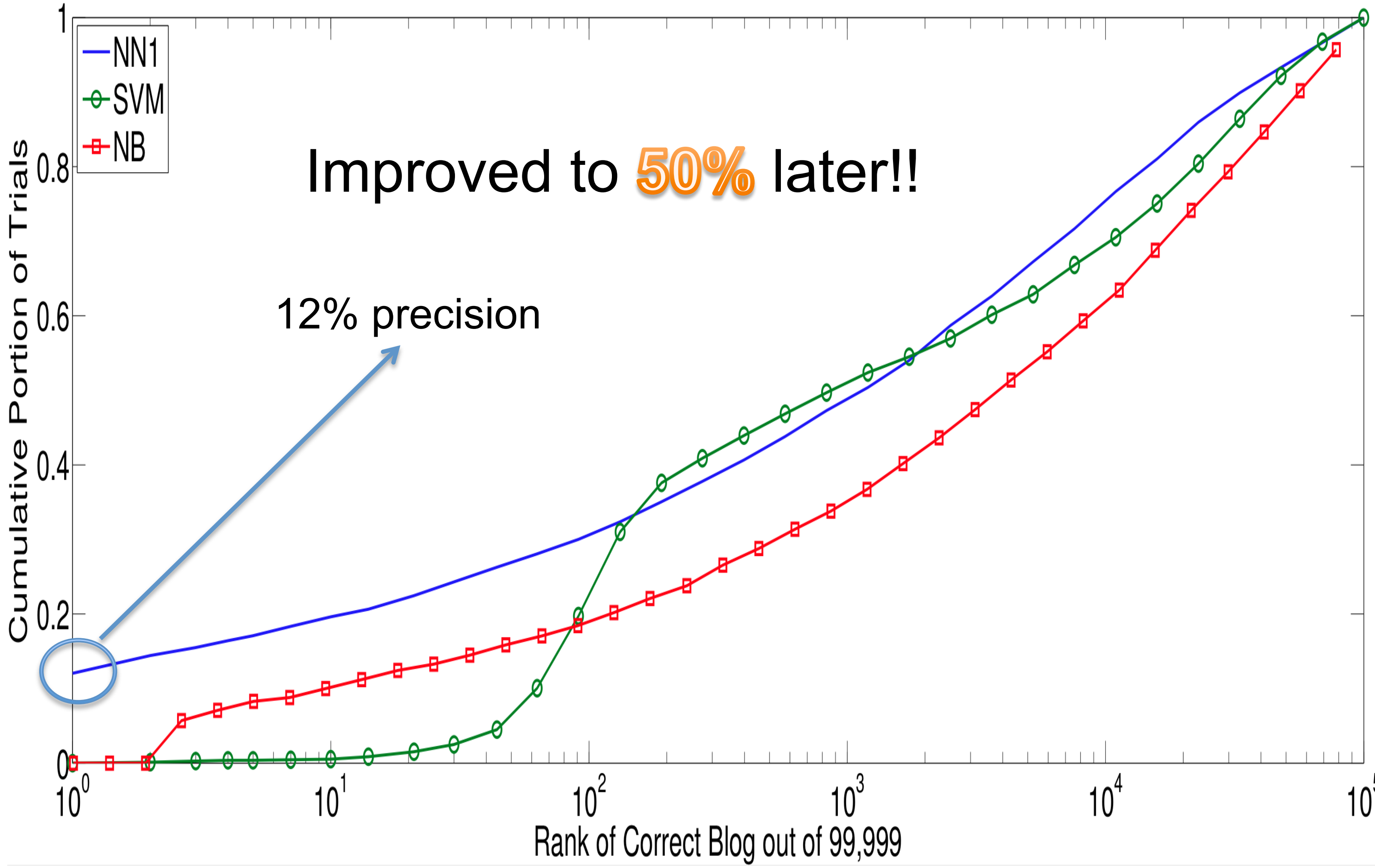
- <u>Post-to-blog experiment</u>
  - Identifying a single or a few anonymous posts.
  - Test posts: random sample a few (e.g., 3) posts from each blog
- <u>Blog-to-blog experiment</u>
  - Identifying an entire blog
  - Test blogs: blogs crawled from URLs specified in Google profiles belong to the same author.

Post-to-blog



Three test posts (roughly 900 words) for each blog.

Blog-to-blog

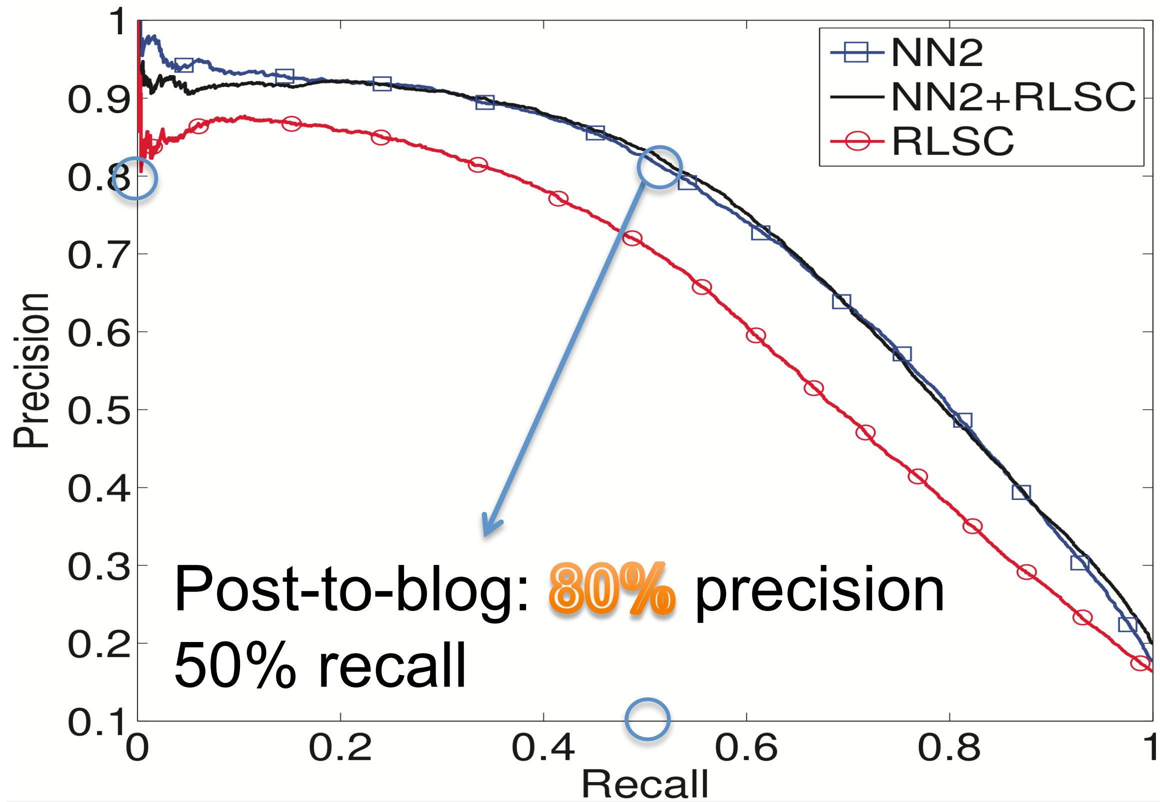


Improved to 50% later!!
12% precision
NN1
SVM
NB
Cumulative Portion of Trials
Rank of Correct Blog out of 99,999
0
0.2
0.4
0.6
0.8
1
10^0
10^1
10^2
10^3
10^4
10^5

Confidence estimation

- Mapping input/output pair of classifier to real values
- Gap Statistics
  - Similarity or distance difference between the best and second best match
- Output the prediction when 'gap' is bigger than some threshold

Confidence estimation



Blog-to-blog: 50% precision. 50% recall

Experiments Summary

- Post-to-blog
  - Best classifier: NN + RLSC.
  - Three test posts, exact match: 20% precision
  - More training/test data, exact match: 40-50%
  - Confidence estimation: 80% precision. 50% recall
- Blog-to-blog
  - Exact match: 12%
  - Confidence estimation: 50% precision. 50% recall

Conclusion

- We identified issues introduced by large scale author identification
- We introduced/discussed strategies to address them
- Large-scale author identification is possible!

- People should be informed
- Be careful when you post sensitive content

Future work

- Better understand what makes authors more/ less fingerprintable
- Design better classifiers
- Automatically transform writing style while preserving document semantics

Thanks!