



Active Authentication From Web Browsing Behavior

US Naval Research Laboratory

PI: **Myriam Abramson, Code 5584**

Myriam.abramson@nrl.navy.mil

Smart Card Alliance Conference
DARPA AAUTH Track
10/30/14



Active Authentication Performer Overview and Status

Naval Research Laboratory (funded by NRL)



PROGRAM OVERVIEW AND STATUS

BIOMETRIC: Identification of users through Web browsing behavior. Develop the theoretical foundations and supporting algorithms for the detection, tracking and prediction of Web browsing behavior using information available from the address line of the browser.

- The biometric is the activity patterns that can be captured in the browser including the timing of clicks, the type of page visited, the length of a session, the revisit rate, etc.

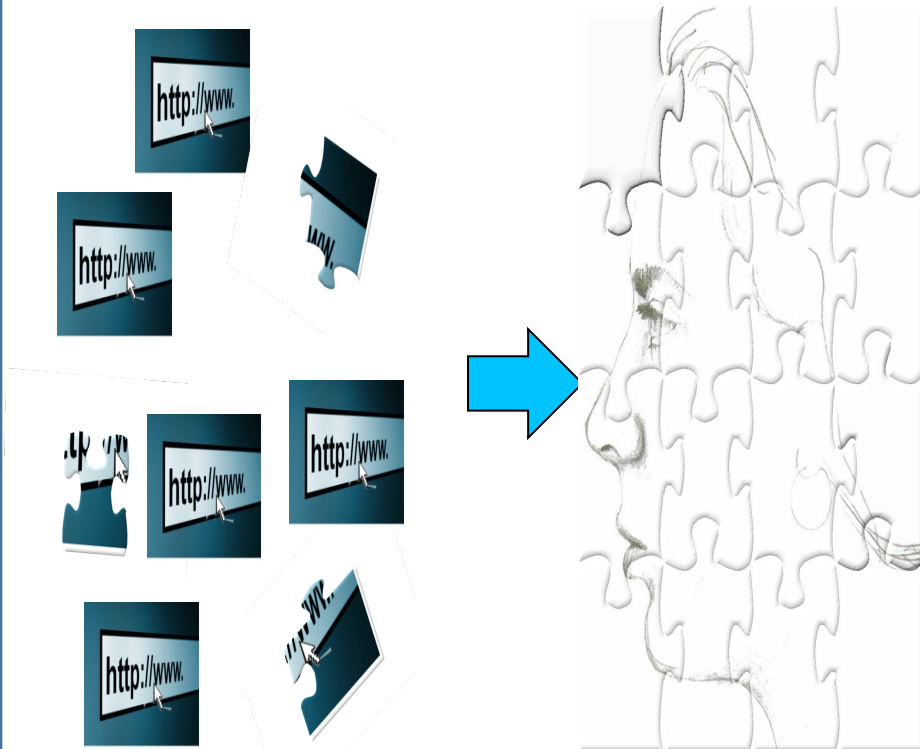
Key Objectives

- Identify, extract and analyze features of Web behavior collected in a user study.
- Investigate structured prediction methods to authenticate users based on their Web browsing behavior.
- Develop a genre palette to categorize webpages for authentication and identification purposes.
- Analyze a large clickstream dataset obtained from comScore, Inc.

Status

- Browser extensions for tracking and monitoring user Web behavior completed and deployed in an ongoing user study.
- Completed detailed analysis of initial user study dataset and identified key features of Web browsing behavior.
- Completed user authentication approach using ensemble of one-class SVMs and random subspace method with best FRR: 11%; and best FAR: 7% ; avg FRR: 17% and avg FAR: 18% for 14 volunteers.
- Completed user authentication of spatio-temporal models with structured prediction methods; avg EER: 15% for 14 volunteers.

Behavioral Web Analytics



Team Members

Principal Investigator: Myriam Abramson, Code 5584

- Aaron Cohen, Code 5545

Outline

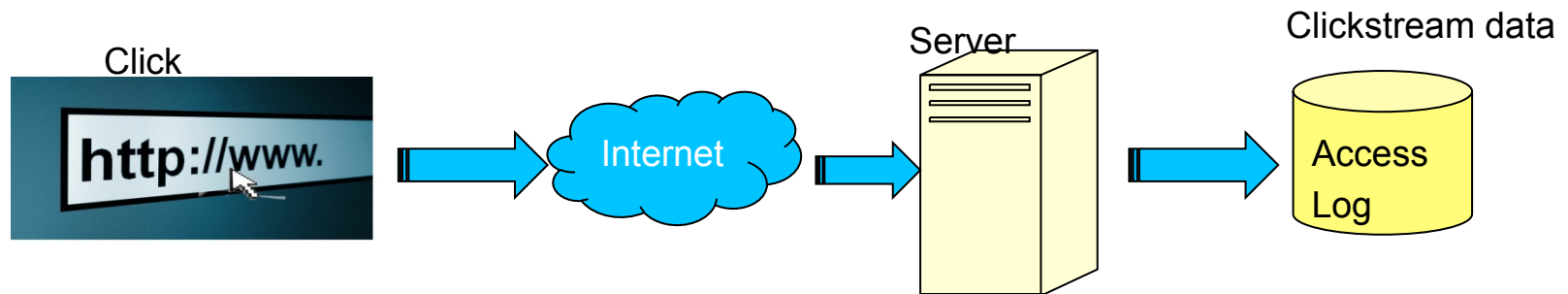


- Objective
- Datasets
- Web behavior features
- BOW approach: Ensemble of One-class SVMs
- Sequence approach: Conditional Random Fields
- Future work

Clickstream data



Clickstream data: UserId, Time, URL visited, browsing agent



Human Subject Research

A screenshot of the Behavioral Web Analytics website. The page has a blue header with the title "Behavioral Web Analytics" and a "Login" link. Below the header is a navigation bar with "Home" and "Upload" tabs. The main content area is divided into two columns. The left column contains sections for "About this Project", "Using the Firefox Extension", and "To install:" followed by a numbered list of steps. The right column contains sections for "Download Firefox Extension" and "Download Chrome Extension", each with a button to download the respective extension file.

Behavioral Web Analytics [Login](#)

Home Upload

About this Project

The objective of this research is to develop the necessary theoretical foundation and supporting algorithms for the detection, tracking and prediction of Web browsing behavior. Toward this end, a Firefox extension has been developed to capture clickstream data within the browser.

Using the Firefox Extension

To install:

1. Download the extension at <https://bwa.nrl.navy.mil> by clicking on the button in the box entitled "Download Firefox Extension".
2. Save the extension on your desktop and drag it into an opened Firefox window to install or open it from the File menu in the toolbar.
3. Install now and restart the browser.
4. A prompt for subject id should come up.
5. Enter whatever you want for subject id (hyphens and spaces will be ignored) or take the default number.

To verify that the extension is working properly:

Under Help -> Troubleshooting information, click on the button "Open Containing Folder" for your Profile directory. You should see your clickstream file under subjectid-clickstream.txt and you can verify with an editor (not Word!) that the URL of the pages you browsed should be written there.

To upload your data:

There should be a Webtracker menu under Tools -> Webtracker. Click on "Directed upload". You should get an alert saying that the upload was successful or an error message.

Download Firefox Extension

Download the most recent version of the extension here.

[webtracker-8.xpi](#)

Download Chrome Extension

Download the most recent version of the extension here.

[webtracker-3.crx](#)

16 volunteers!

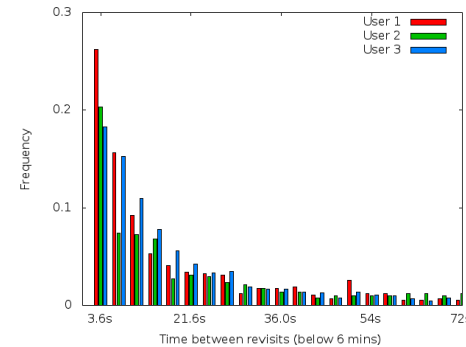
Web Behavior Features



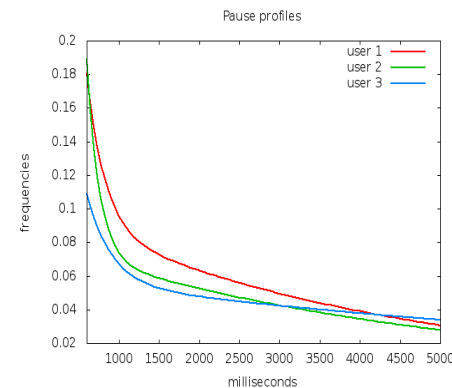
Sessions: series of consecutive clicks delimited by pauses of 30 mins or longer

- **Global session** features
 - Session duration
 - Session length
 - Day-of-week
 - Time-of-day
 - Number of unique domain names
- **Time-variant** distributions
 - Time-between-revisit
 - Pause
 - Burstiness
- **Genre** distribution¹

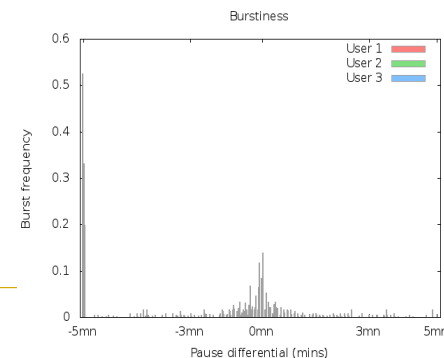
Power law: $y=x^{-\tau}$



Time-between-revisit



Pauses

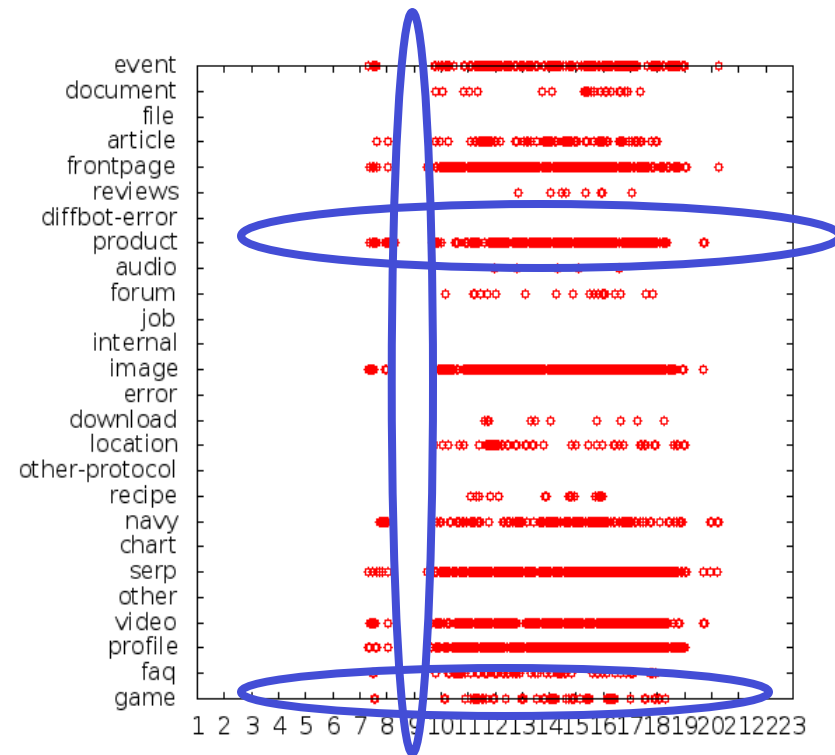
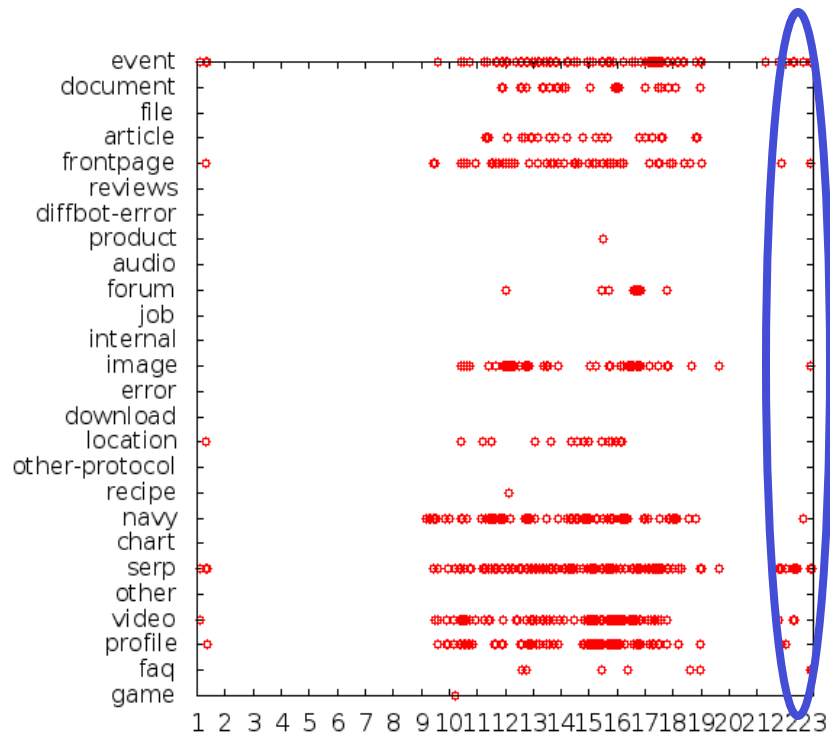


Burstiness

¹ <http://www.diffbot.com>

Associative Patterns of Web Browsing Behavior

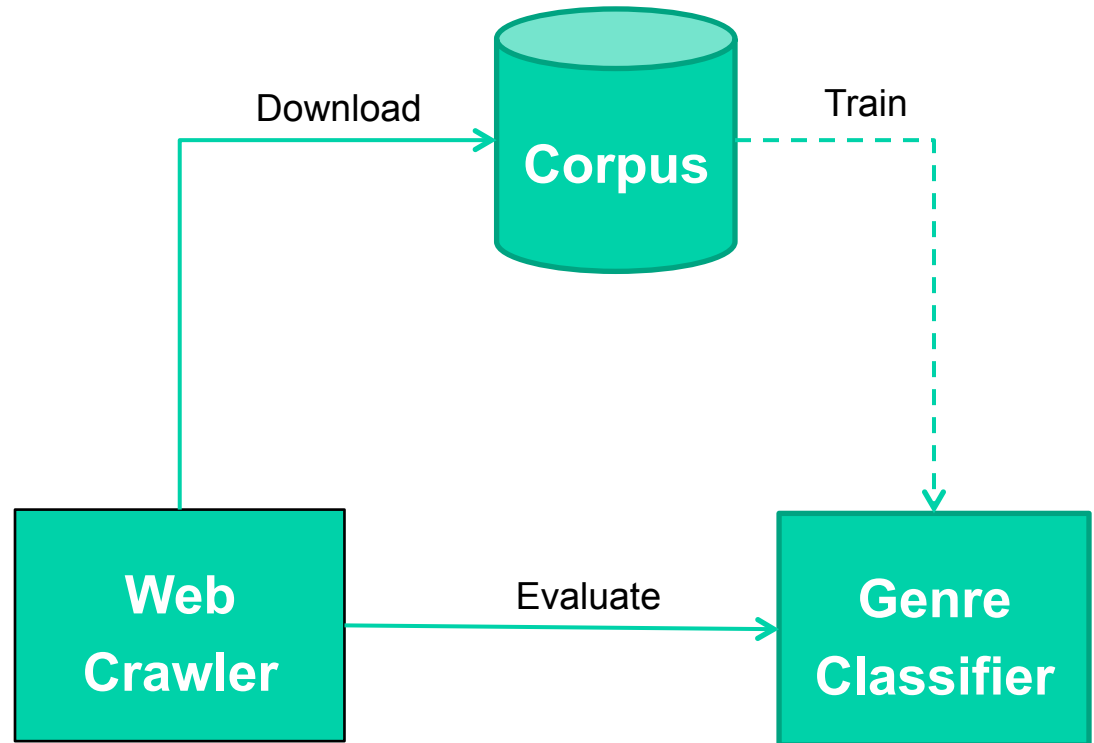
(Abramson et al., AAAI Fall Symp)



Genre Classifier

(with Aaron Cohen, Code 5545)

- Apache Nutch
- Cascalog
- Naïve Bayes¹

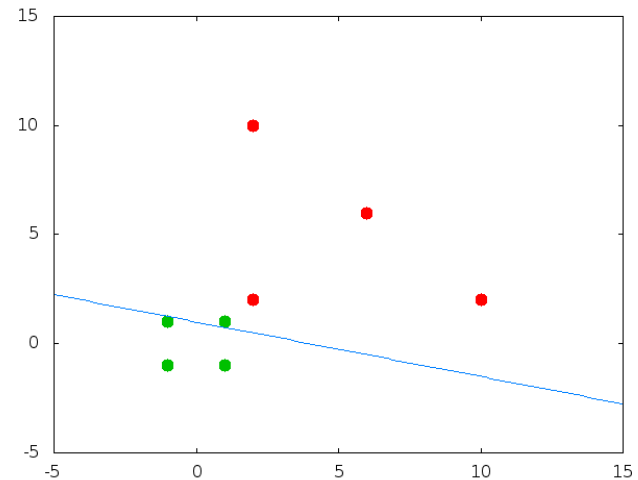


¹ What's in a URL? Genre Classification from URLs (Abramson, et. al, AAAI ITWP, 2012)

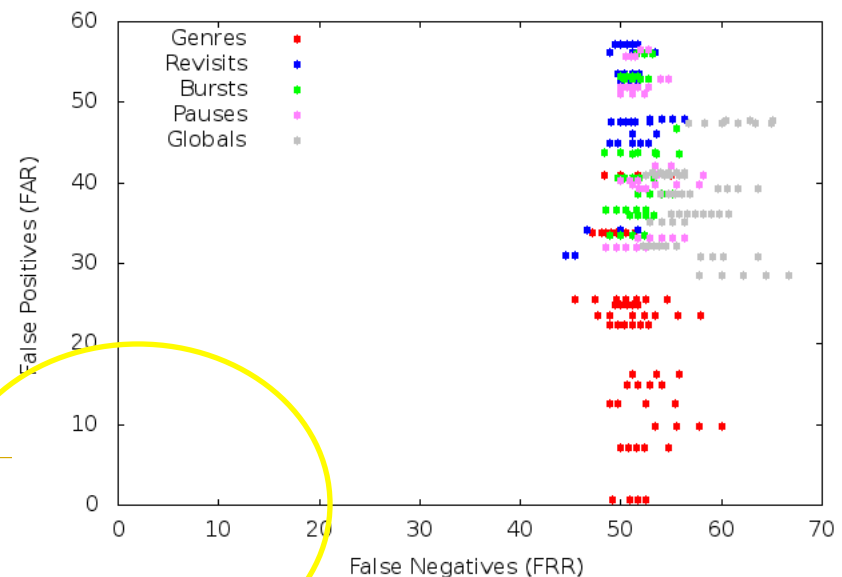
BOW Approach: One-class SVMs



- Unsupervised learning problem
- Like clustering but solves a discriminative problem (self or not self)
- Moves the data to a high-dimensional space with a kernel (e.g. Gaussian Kernel)
- LibSVM: Takes the origin as the only support vector from the complement class
- Authentication Metric: false rejection rate (FRR) and false acceptance rate (FAR)



Evaluation of Web browsing features

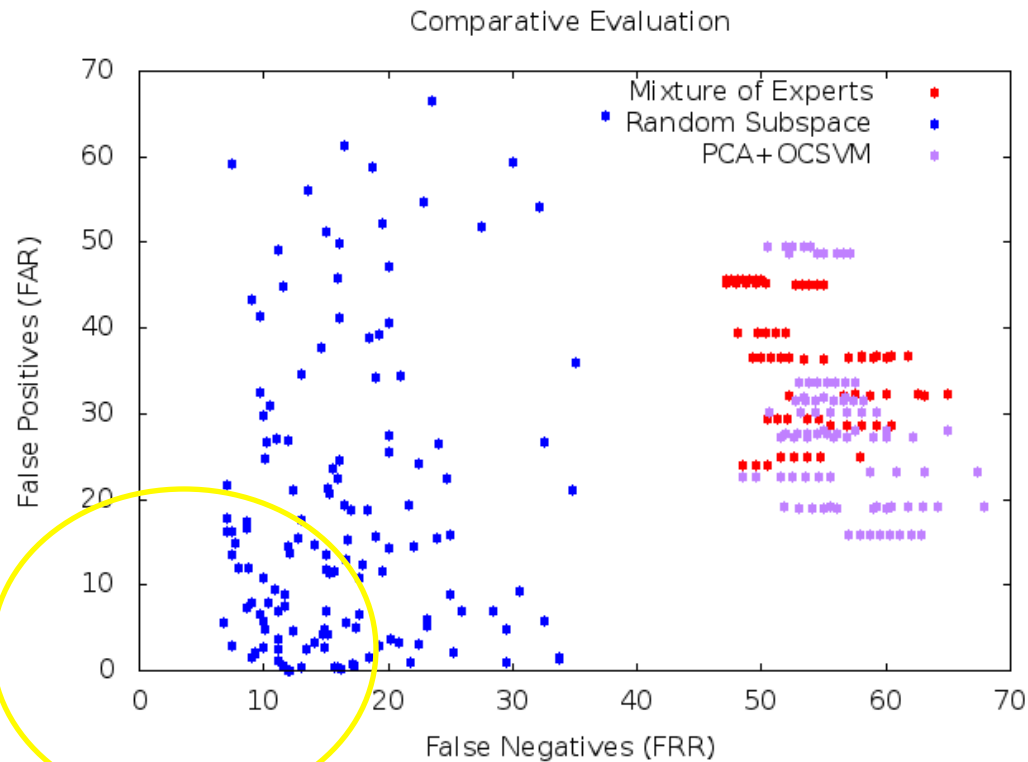


Ensemble Learning: Random Subspace (Abramson, et al., FLAIRS-26)



- Varies the set of features of an ensemble of learners (one-class SVMs) for diversity
 - Pool of learners with different feature sets
 - Select subset of learners with weighted sampling on internal 2-fold cross-validation
- Weighted vote
- Findings:
 - No best feature(s) across **all** volunteers (a profile-based approach should work)
 - Shorter time spans with high resolution are better discriminator

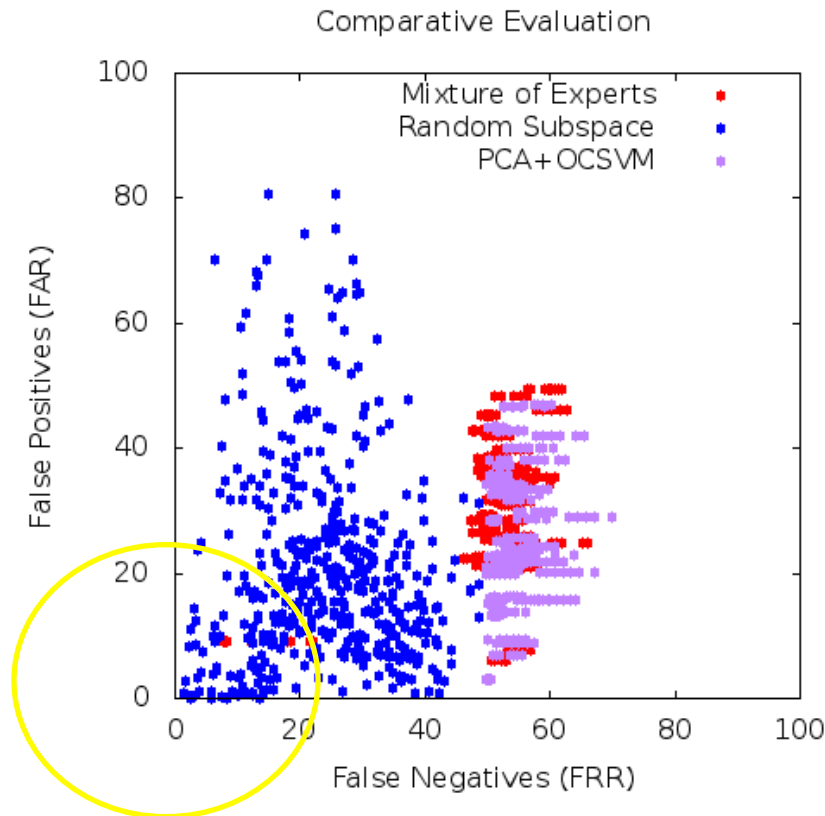
Empirical Results (1499 sessions/14 users/1 month)



	Ensemble/Random Subspace	
	FRR	FAR
AVG	17%	18%

Empirical results: comScore

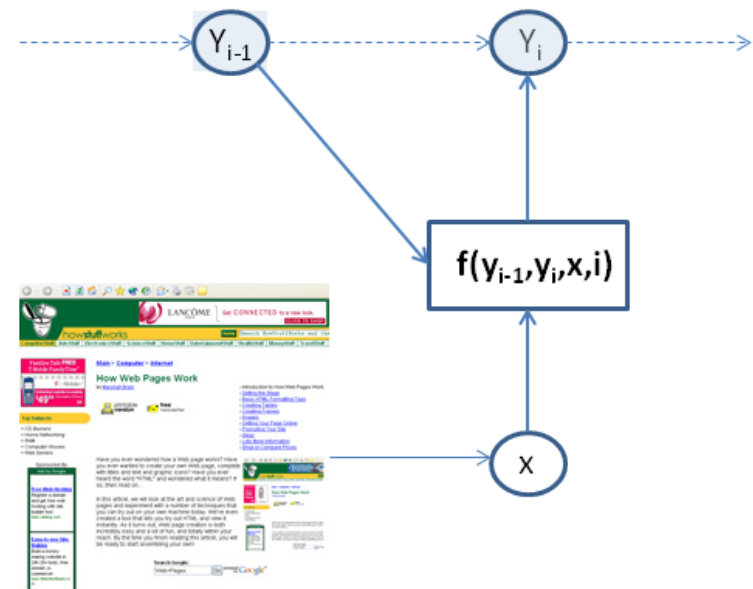
(9567 sessions/47 users/2 months)



	Ensemble/Random Subspace	
	FRR	FAR
AVG	24%	20.6%

Sequence Approach: Conditional Random Fields (CRFs)

- Temporal structured prediction
 - Does order of websites visited matter?
- Linear-chain CRFs:
 - Observations: TOD, DOW, 500 URL ngrams
 - Hidden Nodes: Genre
- Learn a user profile characterized by weights of parameters
 - Stochastic Gradient+Viterbi
 - Structured Perceptron+Viterbi
 - Label Maximization Likelihood (Mallet¹)

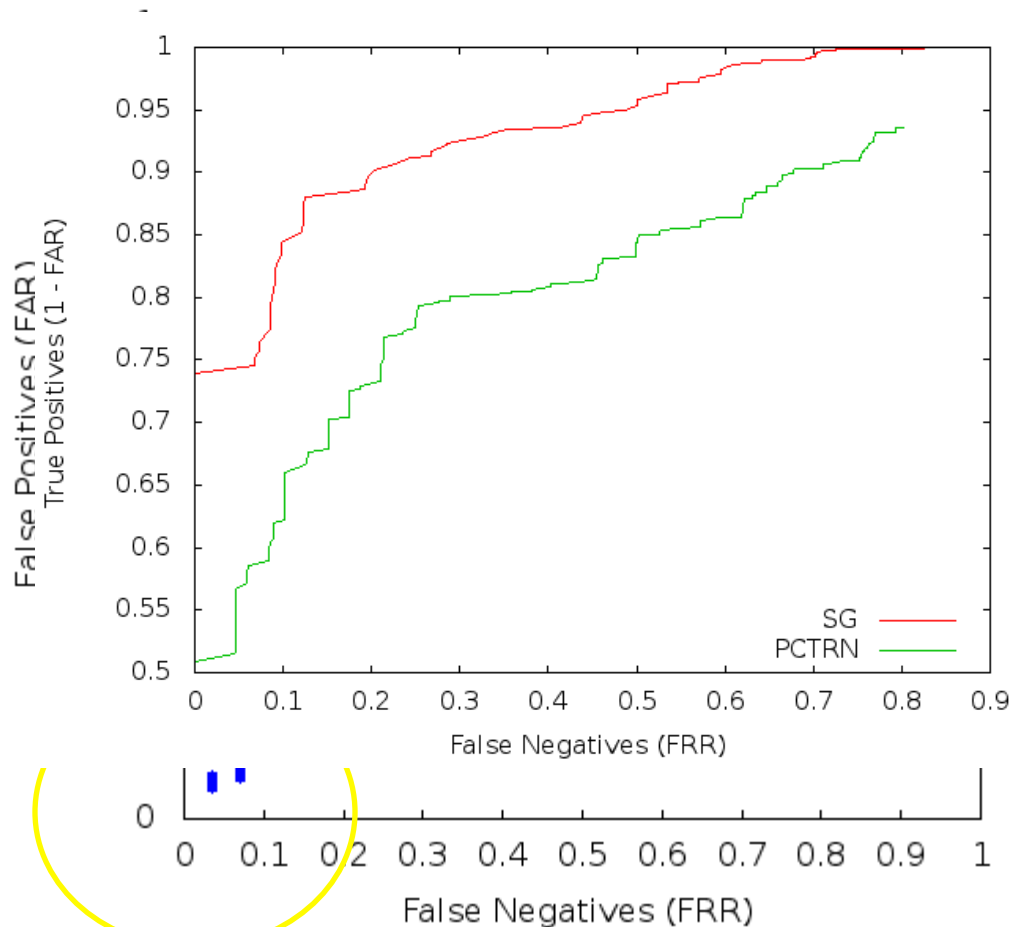


Feature Function

$$p(y_{1:i} | x, \theta, i) = \frac{\exp \sum_{j=1}^n \theta_j f(y_{1:i}, x, i)}{\sum_{y'_{1:i}} \exp \sum_{j=1}^n \theta_j f(y'_{1:i}, x, i)}$$

¹ <http://mallet.cs.umass.edu>

Name that volunteer in 5 clicks! (1147 sessions, 14 users)¹



	SG/Viterbi	
	FRR	FAR
AVG	17%	17%
Best	11%	11%

¹Learning Temporal User Profiles from Web Browsing Behavior (live), M. Abramson, SocialCom 2014

Summary



- Noise is an issue
- Authentication from Web browsing can be done in very few clicks!
- Label accuracy does not always translate into model accuracy.

Future Work



- Scale up
- Robustness of predictive analytics
 - Temporal drift
 - Noise
- Activity patterns
- Generalize to other problems

