Web to World

Predicting Transitions from Self-Diagnosis to the Pursuit of Local Medical Assistance in Web Search

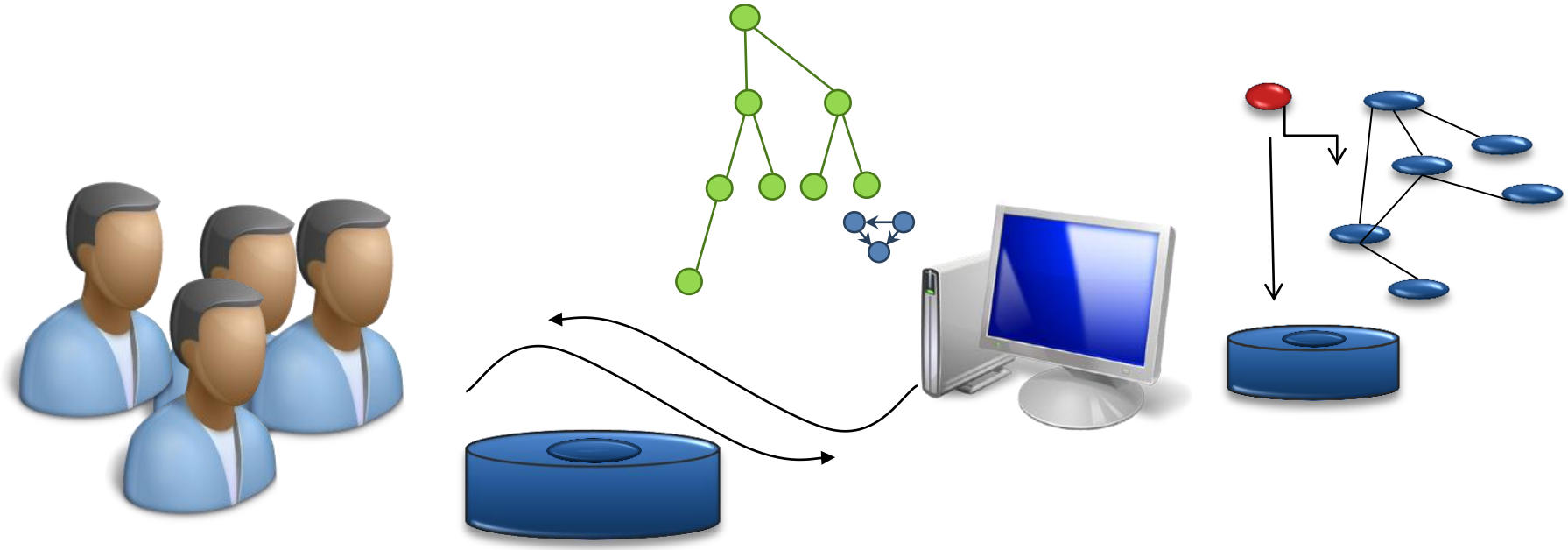
Ryen White, PhD
Eric Horvitz, MD PhD

AMIA
November 2010

Microsoft Research

Pursuit of Insights about Consumer Experiences with Health Search

- Mining insights from large-scale logs
  - Query sequences & page accesses
  - Content distribution & dynamics
  - Insights, predictive models, services



Prior Study: Escalation of Concerns

- Large-scale crawl & log analysis, survey (TOIS 2009)
- Transition from common symptoms to rare diseases

  e.g., {headache, nausea, dizziness} → rare illness

  - <u>Conclusions</u>

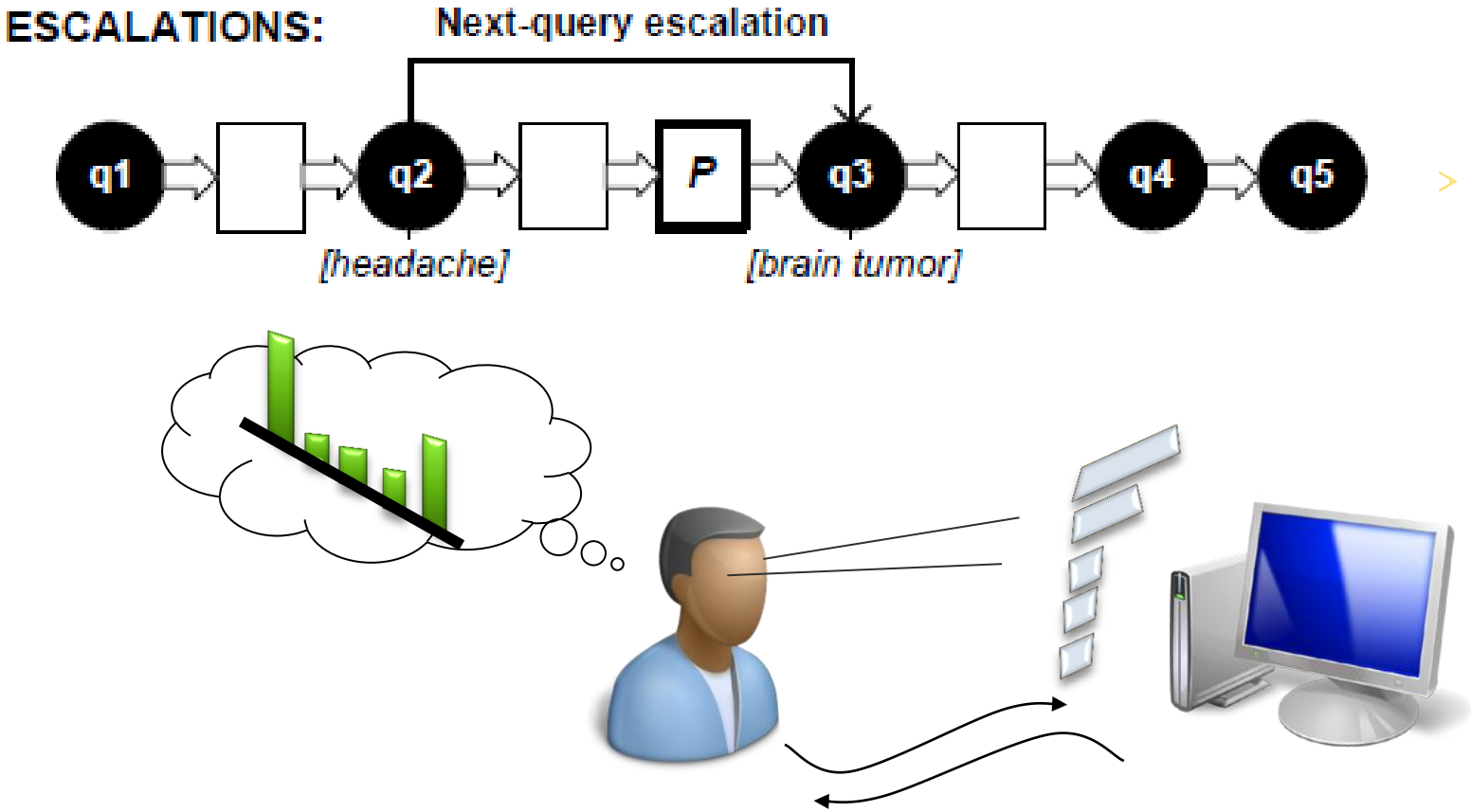
    - Escalations of concerns widespread

    - Web suffers from & amplifies biases of judgment

      - *Base-rate neglect*
      - *Availability bias*

Prior Study: Predicting Escalation

- Predict transition from common symptoms to rare illness *based on features* of pages being viewed (SIGIR 2010)



New Work: Influence of Web on Seeking Healthcare Professionals

- Web search → more engagement with healthcare system (AMIA 2009)
  - Survey of Microsoft employees (n=515):

"Web content put you over the threshold for scheduling an appointment with a health professional, when you would likely have not sought professional medical attention if you had not reviewed Web content."

→ 23.7% Yes!

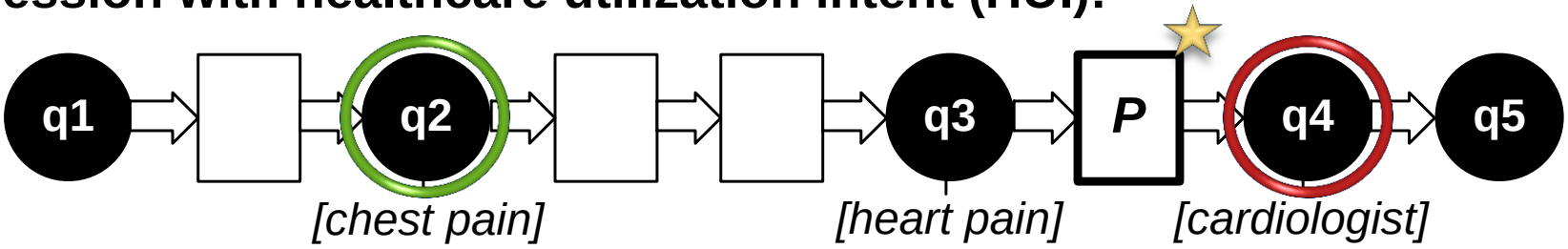
Web to World

- Predict pursuit of in-world healthcare resources: *Healthcare Utilization Intention* (HUI)
  - Querying for information on proximal physicians, specialists, healthcare centers
  - e.g., *"neurologist in seattle, wa"*, *"evergreen hospital"*, *"urgent care clinic"*
- Automated detection:
  - Appropriate medical specialty for the symptom (e.g., *neurologist* for symptom: muscle twitches);
  - medical resource (e.g., *hospital*, *physician*)
  - five-digit US zipcode, US city & state name pair (e.g., *Redmond, Washington*)

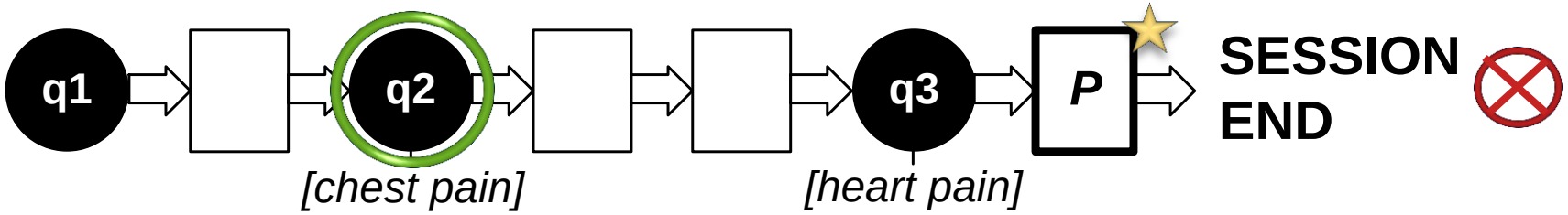
Study of Web to World!

- Prediction of transition to HUI

Session with healthcare utilization intent (HUI):



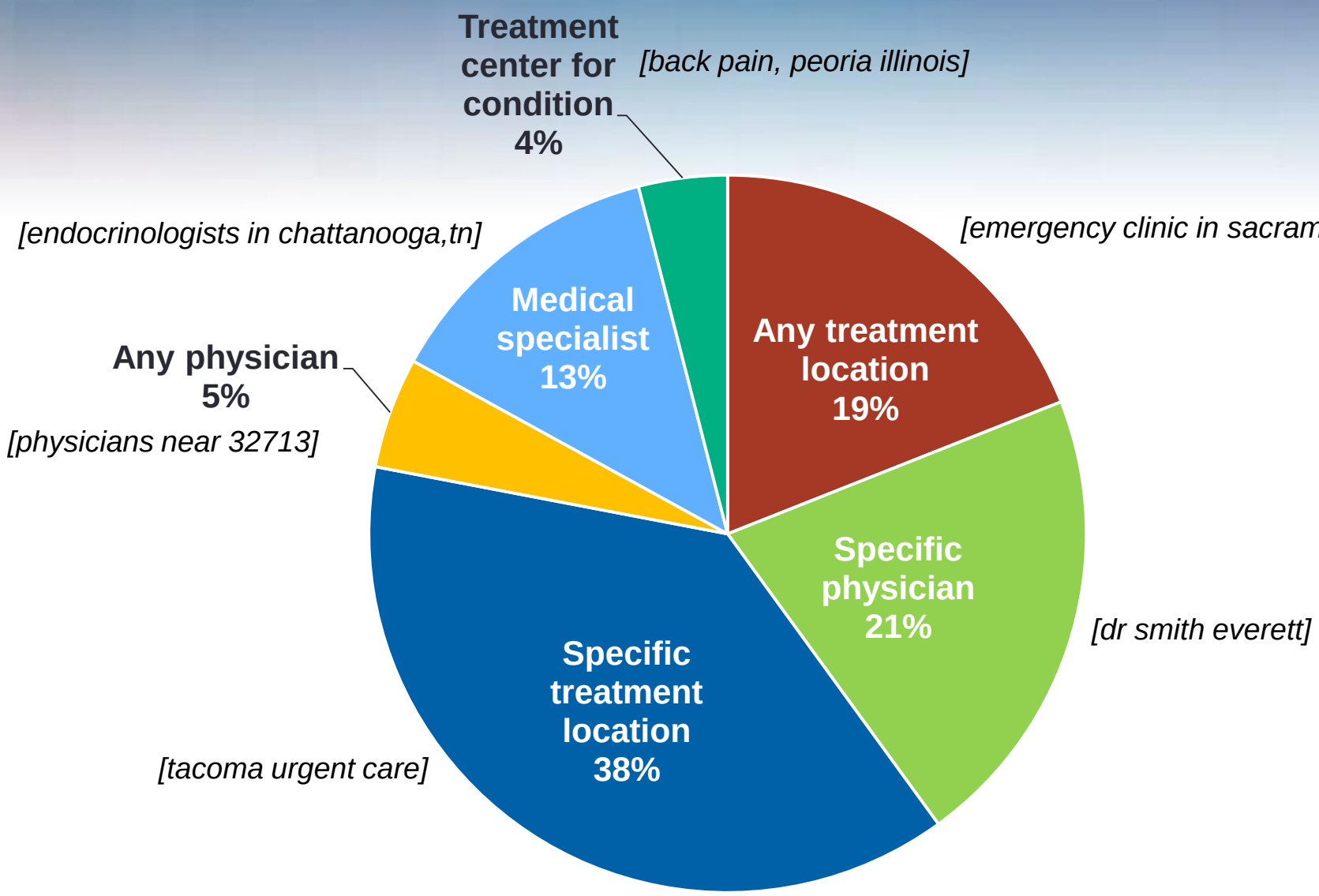
Session without healthcare utilization intent (No HUI):



Methods

- Analysis of Log Data
  - Six mos. anonymized logs from MSN Toolbar (opt in)
  - Extract search sessions using automated tools
  - Sessions: start query and all queries and URLs
  - Symptoms: *chest pain*, *muscle twitches*, *abdominal pain*
  - Automatic extraction of sessions w/ symptom → HUI
  - 700 HUI, 700 no-HUI sessions

Characteristics of Resource Pursuits



Characteristics of Resource Usage

- HUI queries toward end of sessions
  - 36% of sessions, HUI query was last query in session
  - Mean: HUI queries occur 75% of the way through session
- When additional queries follow, search activity is:
  - **Refine** query in pursuit of resource (46%)
  - **Explore** a medical condition (22%)
  - **Compare** different resources (e.g., two specialists) (14%)
  - **Other**
    - Request next search results page (10%)
    - Shift topics (8%)

Predicting Escalations to HUIs

- Prediction task

  Probability that user will next issue an initial HUI query given currently viewing page p.

- Three classes of features
  - **Page:** Structure & content of current page.
  - **Session:** Attributes of search interaction in current session.
  - **User:** Aspects of users' historic medical search interactions from the beginning of our log data to start of current session.

Page Features

FracPageFirstSerious: **Fraction page to first serious illness**

FracPageFirstBenign: **Fraction page to first benign explanation**

NumSeriousInFirstPara: **Number serious illness in first para.**

NumBenignInFirstPara: **Number benign explanations in first para.**

NumNegMod: **Number negative modifiers (e.g., don't have)**

NumPosMod: **Number positive modifiers (e.g., do have)**

NumTestimonials: **Number testimonials (e.g., I was scared)**

UrlTrusted: **Page from trusted source (e.g., medlineplus)?**

TrustedDomain: **Page from trusted domain (e.g., .edu)?**

IsWebForum: **Page from a Web forum?**

HasURACVerification: **Verified by www.urac.org?**

HasHONVerification: **Verified by www.healthonnet.org?**

HasSeekMedicalAdvice: **Recommends medical consult.?**

ForHealthProfessionals: **Content meant for health prof.?**

LengthInWords: **Number of words**

SizeInKB: **Size in kilobytes (text only)**

HasResources: **Mentions external resources (e.g., doctor)?**

Page Features

AdsPresent: **Advertisements present on page?**

NumAdBlocks: **Number of advertising blocks**

SeriousThenBenign: **Serious illness for concern appears on the page before a benign explanation for that symptom?**

NumWordsToSerious: **Number words to first serious illness**

NumWordsToBenign: **Number words to first benign explanation**

NumWordsBetweenSeriousAndBenign: **Number words between first serious illness and first benign explanation**

SeriousInTitle: **Serious illness in page title?**

BenignInTitle: **Benign explanation in page title?**

SeriousInFirstPara: **Serious illness in first paragraph?**

BenignInFirstPara: **Benign explanation in first paragraph?**

SeriousAndBenignInFirstPara: **Serious/benign first para.?**

NoSeriousBenignInFirstPara: **No serious/benign first para.?**

NumSerious: **Number serious illnesses**

NumBenign: **Number benign explanations**

NumGraveConcerns: **Number grave concerns (e.g., fatal)**

Session and User Features

NumQueries: **Number queries**

AvgQueryLength: **Average query length (in tokens)**

NumEscQueries: **Number queries with escalations for concern**

NumNonEscQueries: **Number queries with benign explanation**

NumURLs: **Number (non- search engine result) pages**

AvgDwellTime: **Average dwell time on pages**

TotalDwellTime: **Total dwell time on pages**

AvgConcernSearchesPerDay: **Number concern queries per day**

AvgMedicalSessionsPerDay: **Number medical sessions per day**

NumUniqueSymptoms: **Number unique Merck symptoms**

NumEscalations: **Number previous queries for serious illnesses**

Exploration of Key Features

- Explore page, session, user features

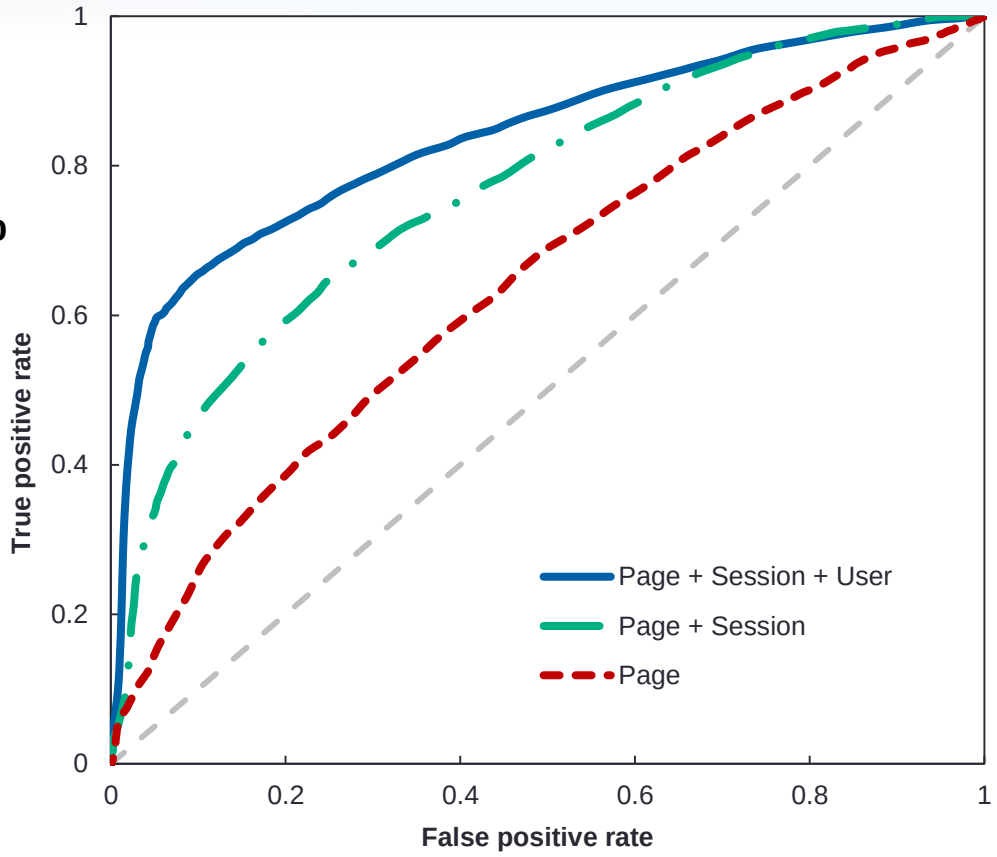
Features	HUI	No HUI
SeriousBeforeBenign (Page)	59%	48%
IsWebForum (Page)	14%	9%
NumQueries (Session)	4.9	2.9
AvgQueryLength (Session)	4.5	4.1
NumUniqueSymptoms (User)	3.6	2.2
NumResourceQueries (User)	5.5	2.0

- *All differences are significant*

Study of Predictive Model

Logistic regression with five-fold cross-validation

- Accuracy:
  - Page features = 59.3%
  - Page + session = 68.9%
  - **Page + session + user = 77.7%**



Prediction Findings

- Inspected feature weights
- Top features by evidential weight, relative to most predictive feature, *AvgDwellTime*:

Feature	Class	Importance
AvgDwellTime	Session	1.00
NumEscalations	User	0.71
HasResources	Page	0.60
NumResourceQueries	User	0.56
NumURLs	Session	0.47

Value of multiple classes of features in building predictive models

Summary

- Web to world: Predicting *Health Utilization Intention* (HUI)
- Predictive models of escalation to HUI given features of a page, session, user
- Characterized resource seeking:
  - Most HUIs are searches for specific locations or physicians
  - Post-initial HUI query, users refine, explore, or compare