Enriching the Web by Modeling Reading Difficulty

Kevyn Collins-Thompson

Associate Professor, University of Michigan

ESAIR 2013: *Exploiting Semantic Annotations in Information Retrieval*
October 28, 2013

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

EECS ECE CSE

Acknowledgements

Joint work with my collaborators:

Paul Bennett, Ryen White, Sue Dumais (*MSR*)

Jin Young Kim (*Microsoft*)

Sebastian de la Chica (*Microsoft*)

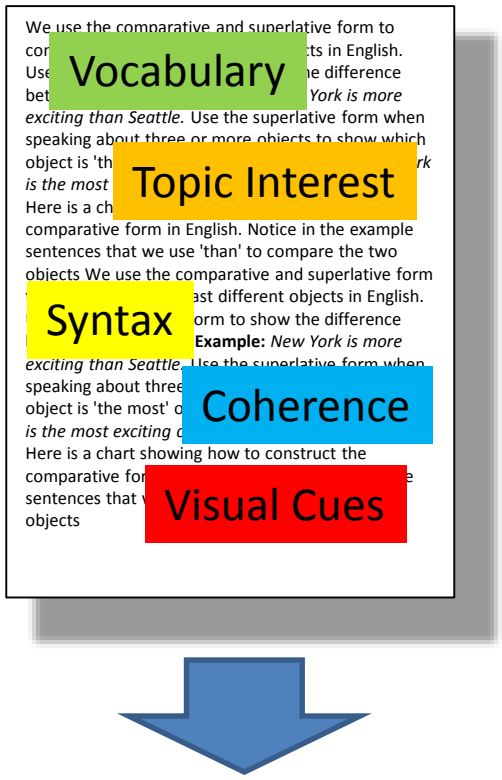
Paul Kidwell (*LLNL*)

Guy Lebanon (*Amazon*)

David Sontag (*NYU*)

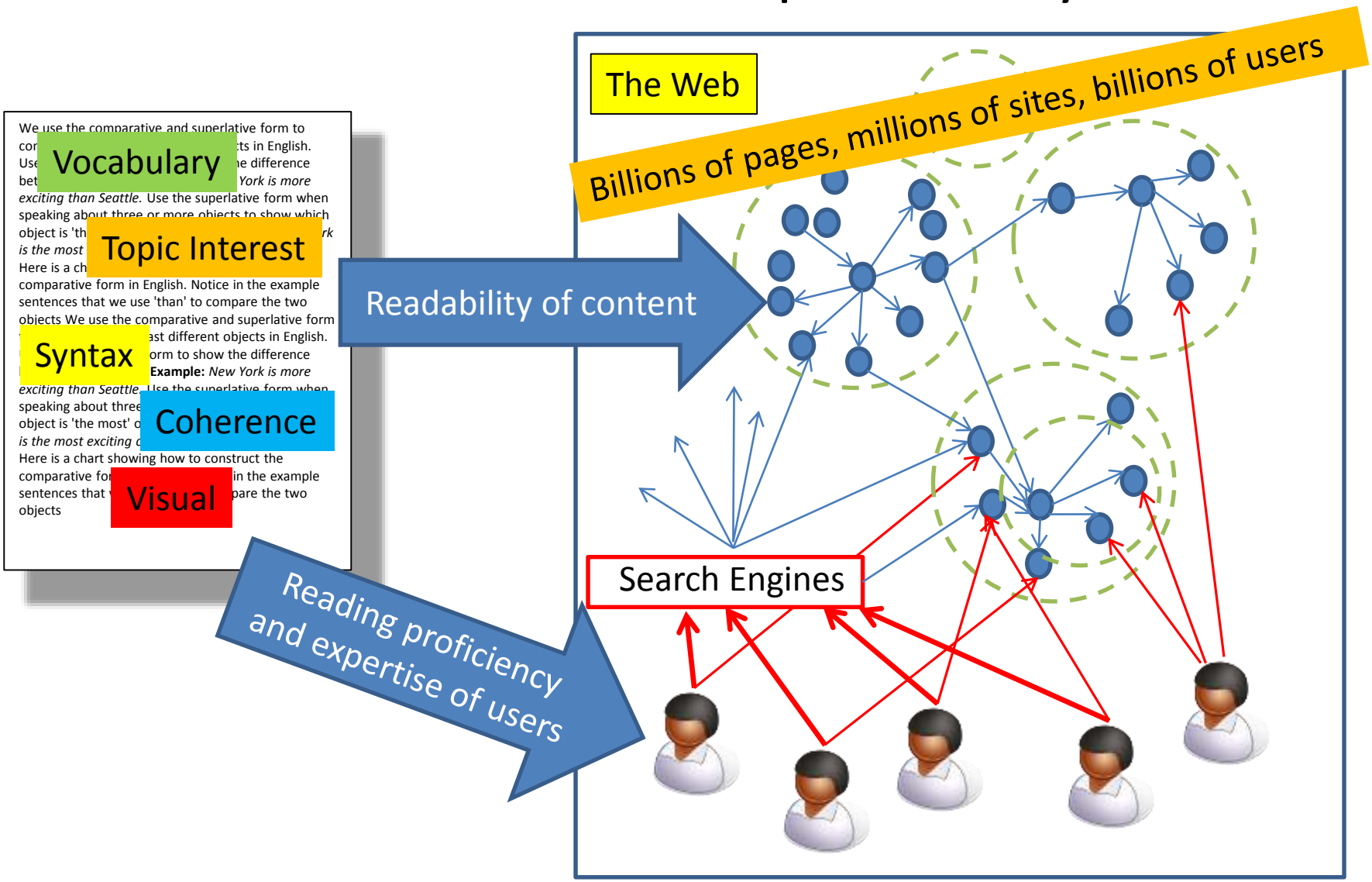
Bringing together readability and the Web … sometimes in unexpected ways

Text Readability Modeling and Prediction

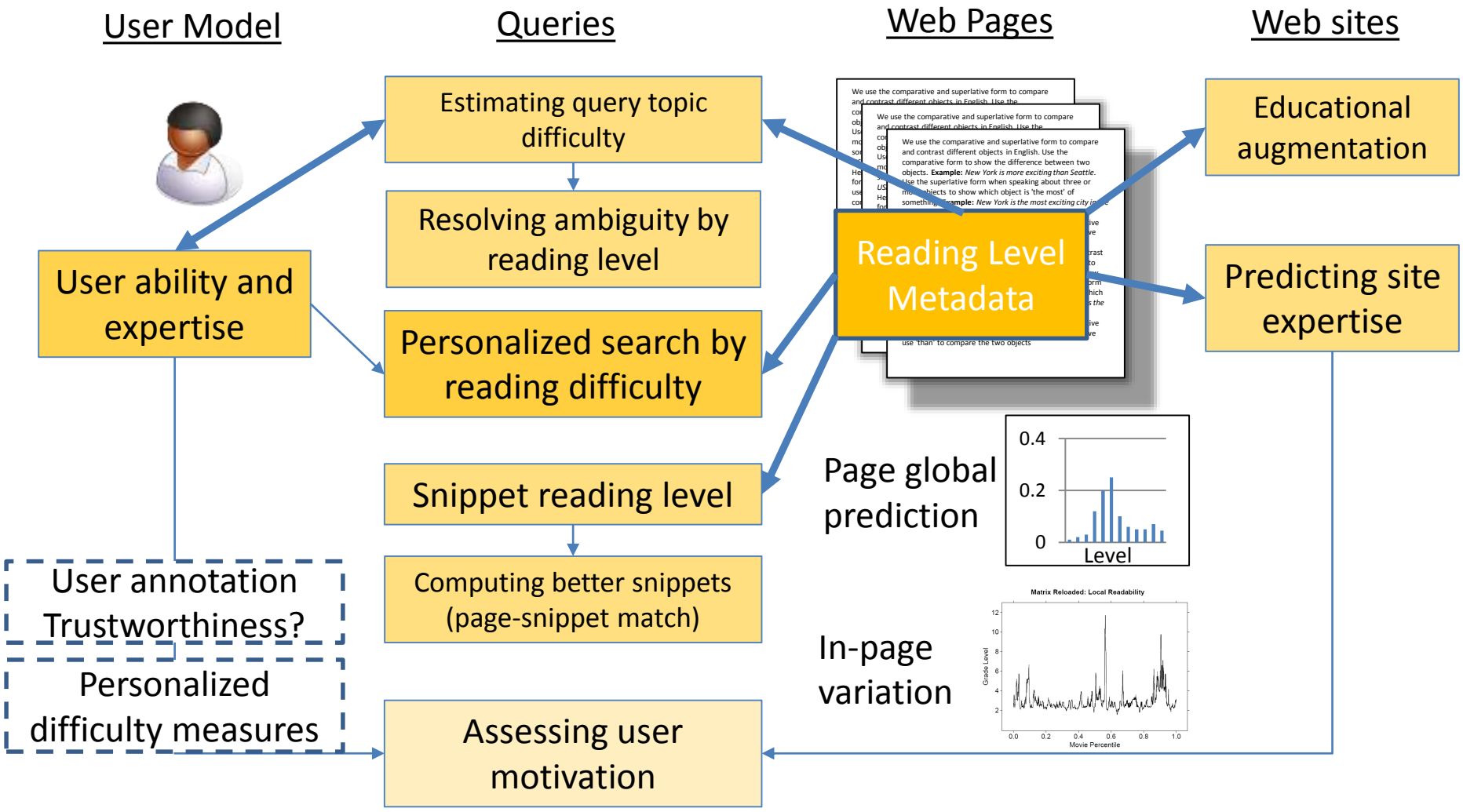


Reading level prediction
Topic prediction

Bringing together readability and the Web … sometimes in unexpected ways



How modeling reading difficulty enriches the Web: Adding reading level metadata to pages leads to novel applications and unexpected insights



Web pages occur at a wide range of reading difficulty levels

Grasshopper Habitat and Grasshopper Diet

Grasshoppers live in fields, meadows and just about anywhere they can find generous amounts of food to eat. A grasshopper has a hard shell and a full grown grasshopper is about one and a half inches, being so small you would not think they would eat much - but you would be so wrong - they eat lots and lots - an average grasshopper can eat 16 time its own weight.

The grasshoppers favourite foods are grasses, leaves and cereal crops. One particular grasshopper - the Shorthorn grasshopper only eats plants, but it can go berserk and eat every plant in sight - makes you wander where they put it all.

Grasshopper Behaviour

Query [insect diet]: Lower difficulty

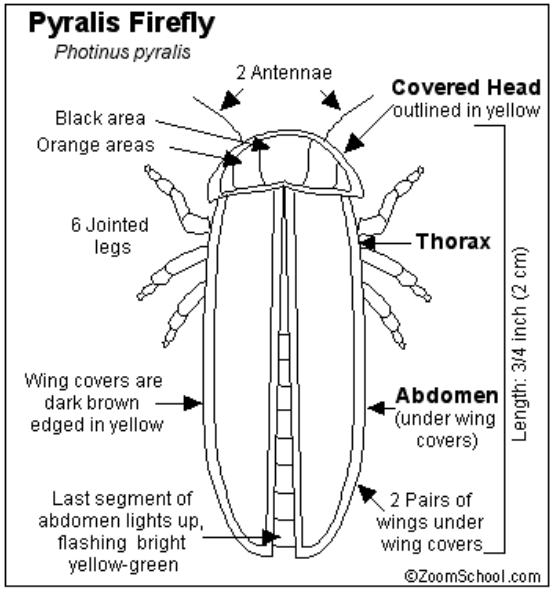
Medium difficulty [insect diet]

Insect Printouts

Firefly or Lightning Bug

Photinus pyralis

More Printouts



The Pyralis firefly (also known as the lightning bug) is a common firefly in North America. This partly nocturnal, luminescent beetle is the most common firefly in the USA.

The Firefly's Glow: At night, the very end (the last abdominal segment) of the firefly glows a bright yellow-green color. The firefly can control this glowing effect. The brightness of a single firefly is 1/40 of a candle. Fireflies use their glow to attract other fireflies. Males flash about every five seconds; females flash about every two seconds. This firefly is harvested by the biochemical industry for the organic compunds luciferin (which is the chemical the firefly uses for its bioluminescence).

Anatomy: This flying insect is about 0.75 inch (2 cm) long. It is mostly black, with two red spots on the head cover; the wing covers and head covers are lined in yellow. Like all insects, it has a hard exoskeleton, six jointed legs, two antennae, compound eyes, and a body divided into three parts (the head, thorax, and abdomen).

Diet: Both the adults and the larvae are **carnivores** (meat-eaters). They eat other insects (including other fireflies), insect larvae, and snails.

Higher difficulty [insect diet]

R&D

INSECT REARING

RESEARCH and DEVELOPMENT

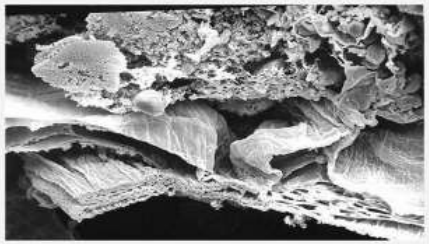
at NCSU

REARING RESEARCH ON DIET DEVELOPMENT AND ESTABLISHING NEW AND IMPROVED REARING SYSTEMS

- **Development of artificial diets and rearing systems for many species of insects has been our specialty.**
- **We use a variety of techniques to help us develop appropriate diets, including analysis of the natural foods, feeding biology of the target insects, bioassays, biological/biochemical testing, and analysis of internal biology.**

Some of our recent and current projects are below:

Here are cactus moth larvae (Cactoblastis cactorum: Lepidoptera: Pyralidae) developing on one of our newest artificial diets developed for USDA, APHIS and Florida Department of Plant Industry.

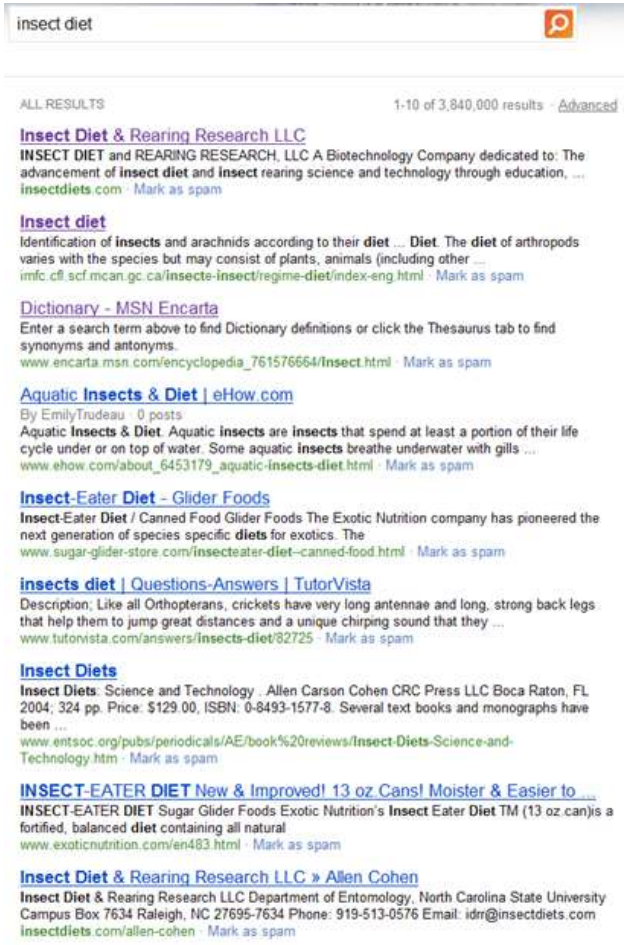


The above electron microscope image shows the peritrophic matrix (PM) of a tobacco budworm fed plant parts from its natural diet: note the multiple layers formed in response to a natural food.

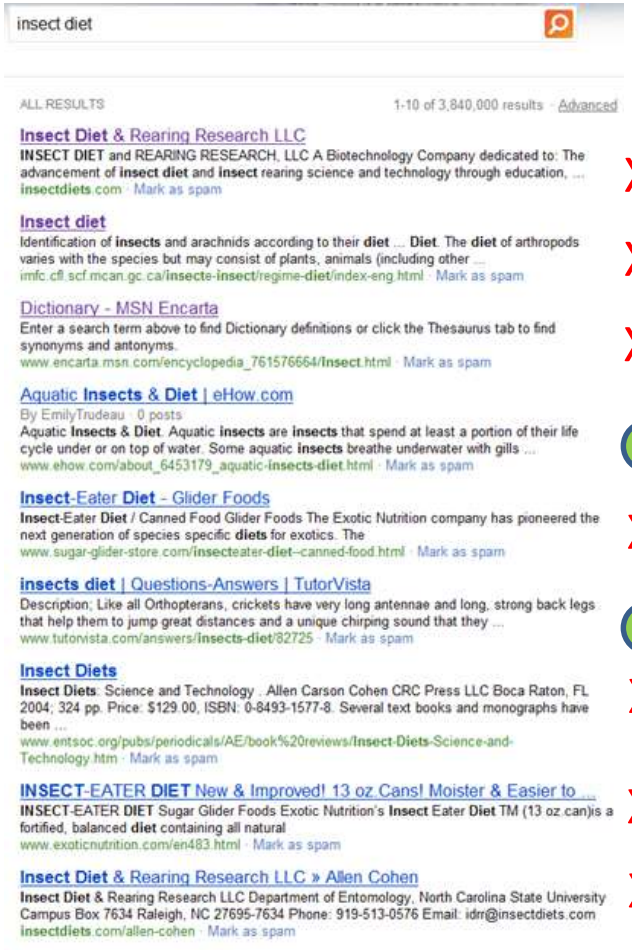
Users also exhibit a wide range of proficiency and expertise

- Students at different grade levels
- Non-native speakers
- General population
  - Large variation in language proficiency
  - Special needs, language deficits
  - Familiarity or expertise in specific topic areas
- Even for a single user there can be broad variation in intent across search queries

Default results for [insect diet]



Relevance as seen by an elementary school student (e.g. age 10)



X Technical

X Technical

X Relevance

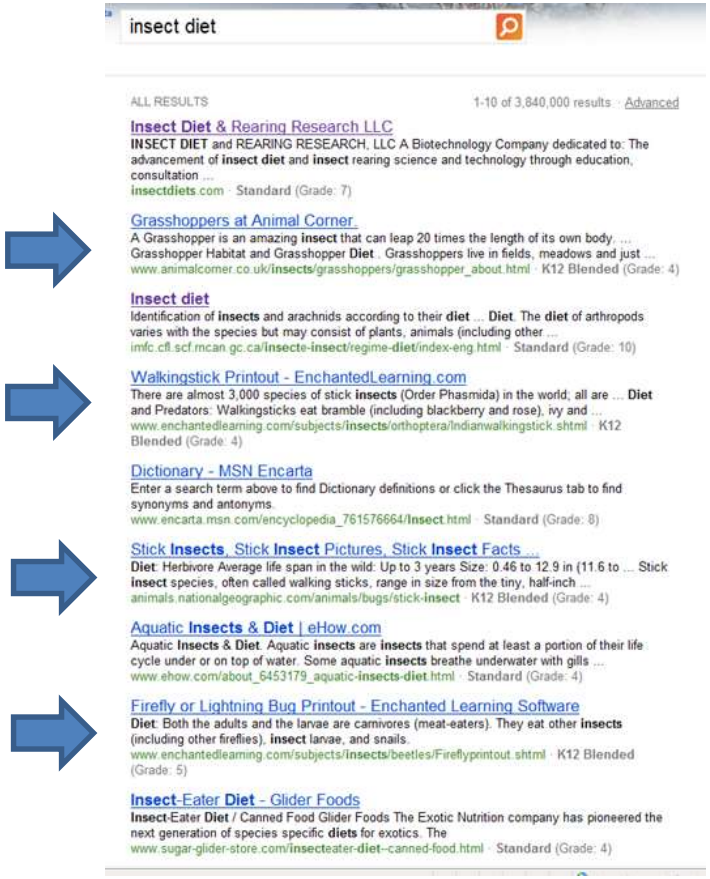
X Relevance

X Technical

X Relevance

X Technical

Blending in lower difficulty results would improve relevance for this user



X Technical

X Technical

X Relevance

X Relevance

Reading difficulty has many factors

- Factors include:
  - Semantics, *e.g. vocabulary*
  - Syntax, *e.g. sentence structure, complexity*
  - Discourse-level structure
  - Reader background and interest in topic
  - Text legibility
  - Supporting illustrations and layout
- Different from parental control, UI issues

Traditional readability measures don't work for Web content

- Flesch-Kincaid (Microsoft Word)

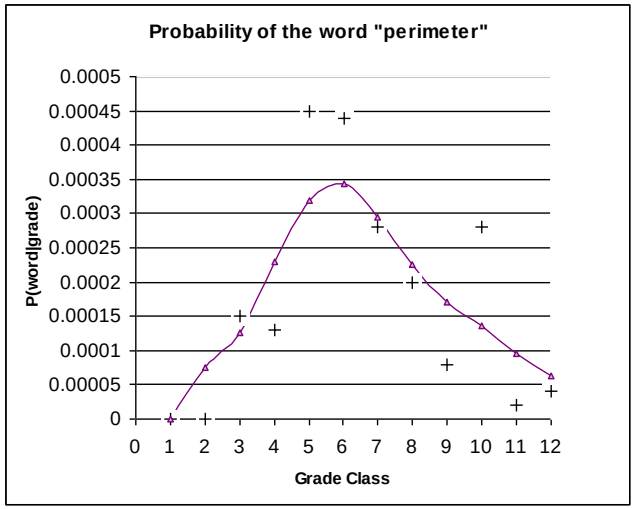
$$RG_{FK} = 0.39 \cdot [Words / Sentence] + 11.8 \cdot [Syllables / Word] - 15.59$$

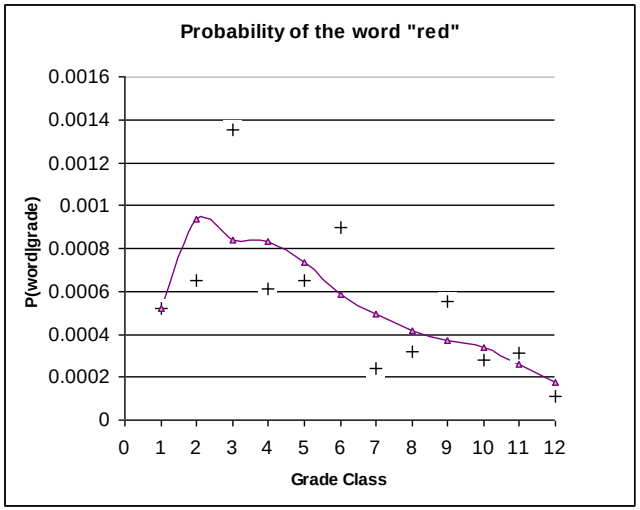
- Problems include:
  - They assume the content has well-formed sentences
  - They are sensitive to noise
  - Input must be at least 100 words long
- Web content is often short, noisy, less structured
  - Page body, titles, snippets, queries, captions, …
- Billions of pages → computational constraints on metadata types
- We focus on *vocabulary-based prediction models* that learn fine-grained models of word usage from labeled texts

Method 1: Mixtures of language models that capture how vocabulary changes with level

[*Collins-Thompson & Callan: HLT 2004*]

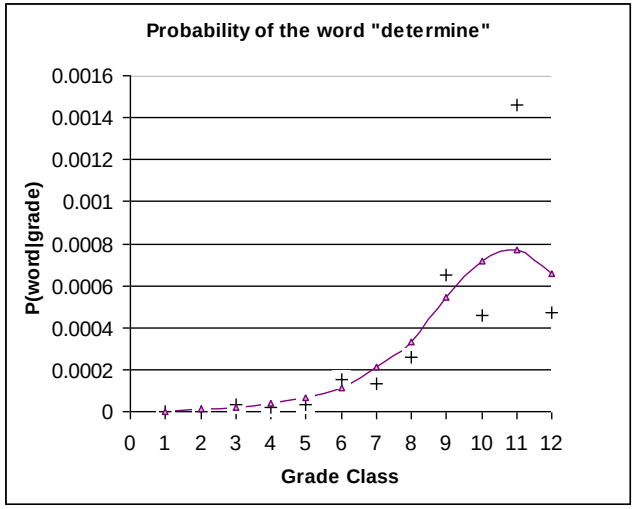
perimeter

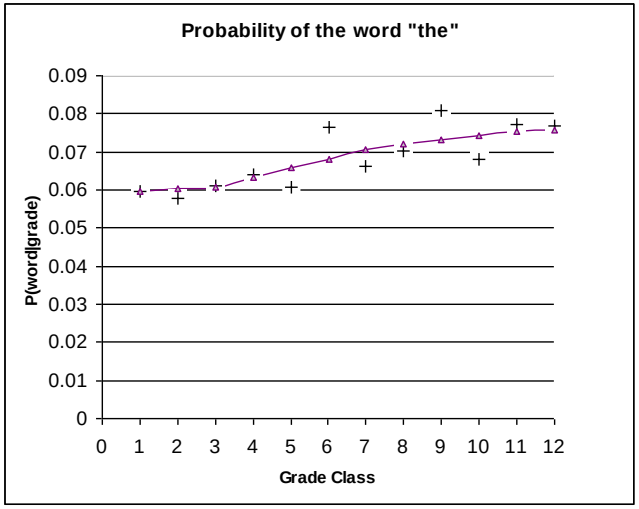




red

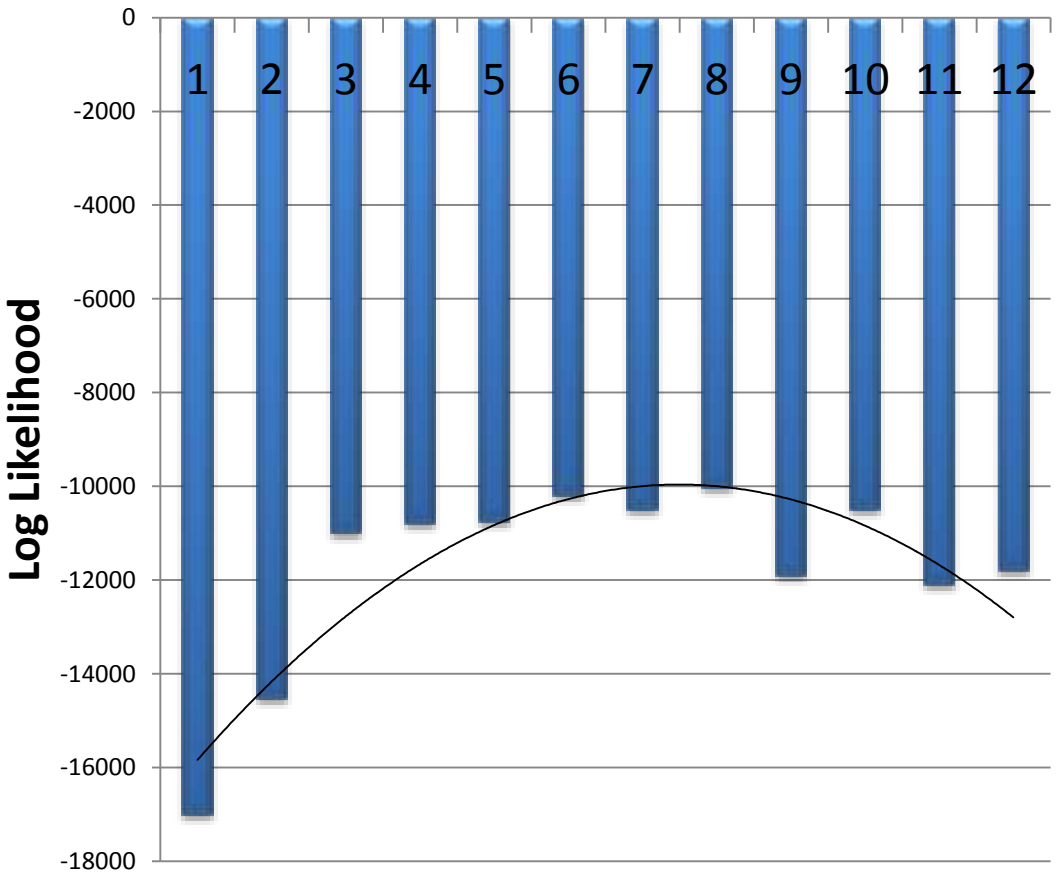
determine





the

Grade level likelihood usually has a well-defined maximum



Grade 8 document: 1500 words

Method 2: Vocabulary-based difficulty measure via word acquisition modeling

[*Kidwell, Lebanon, Collins-Thompson: EMNLP 2009, JASA 2011*]

- Documents can contain high-difficulty words but still be lower grade level
  - e.g. teaching new concepts

- We introduce a statistical model of **(*r, s*) *readability***
  - ***r*** : familiarity threshold for any word

    A word *w* is familiar at a grade if known by at least *r* percent of population at that grade
  - ***s*** : coverage requirement for documents

    A document d is readable at level t if s percent of the words in d are familiar at grade t.

- Estimate word acquisition age Gaussian (μ_w, σ_w) for each word *w* from labeled documents via maximum likelihood
- (*r*, *s*) parameters can be learned automatically or specified to tune the model for different scenarios

We can use these word usage trends to compute feature weights per grade

Grade 1	
grownup	2.485
ram	2.425
planes	2.411
pig	2.356
jimmy	2.324
toad	2.237
shelf	2.192
cover	2.184
spot	2.174
fed	2.164

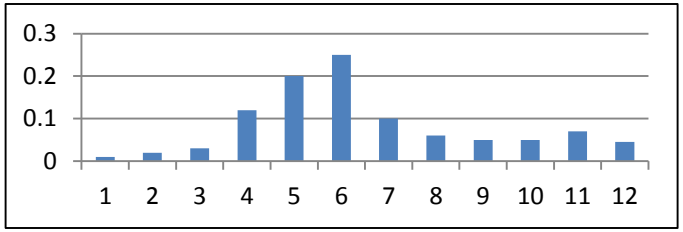
Grade 4	
desert	1.787
crew	1.765
habitat	1.763
butterflies	1.758
rough	1.707
slept	1.659
bowling	1.643
ribs	1.610
grows	1.606
entrance	1.604

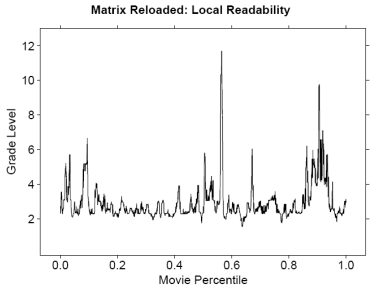
Grade 8	
acidic	1.425
soda	1.425
acid	1.408
typical	1.379
angle	1.362
press	1.318
radio	1.284
flash	1.231
levels	1.229
pain	1.220

Grade 12	
essay	2.441
literary	2.383
technology	2.363
analysis	2.301
fuels	2.296
senior	2.292
analyze	2.279
management	2.269
issues	2.248
tested	2.226

New metadata based on reading level

- Documents:
  - Posterior distribution over levels
  - Distribution statistics:
    - Expected reading difficulty
    - Entropy of level prediction
  - Temporal / positional series
  - Vocabulary models
    - Key technical terms
    - Regions needing augmentation (Text, images, links to sources)



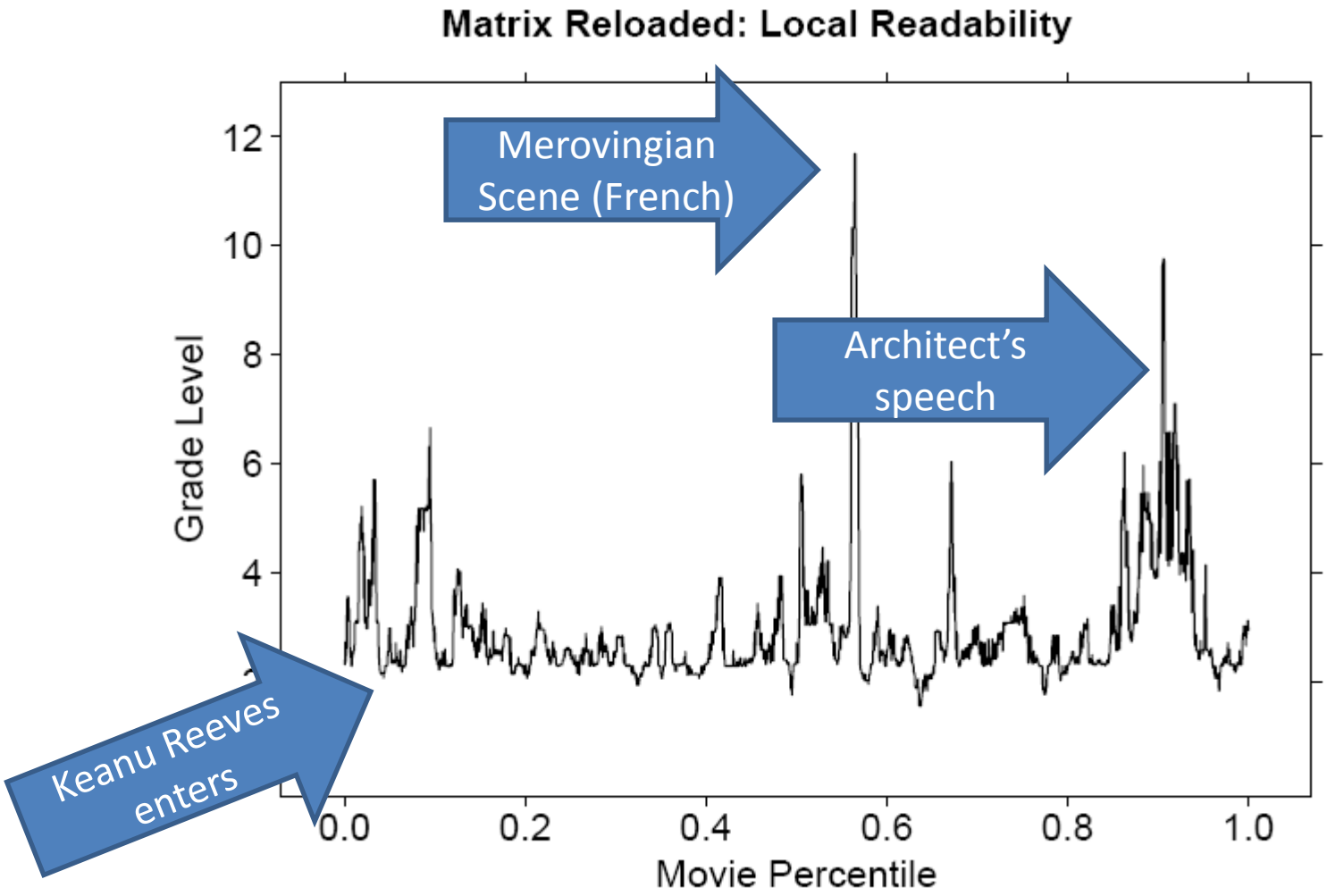


Health article: *Bronchitis*, *efficacy* …

- Web sites:
  - Topic, reading level expectation and entropy across pages
- User profiles:
  - Aggregated statistics of documents and sites based on short- or long-term search/browse behavior

Local readability within a document

Movie dialogue in “The Matrix: Reloaded”



[Kidwell, Lebanon, Collins-Thompson. J. Am. Stats. 2011]

Application: Personalizing Search Results by Reading Level

[*Collins-Thompson et al., CIKM 2011*]

Search engines try to maximize relevance but have traditionally ignored text difficulty

It's not relevant (at least, not immediately)

…if you can't understand it.

A search result should be at the reading level the user wants for that query.

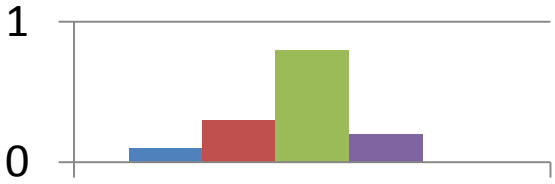
Intent Models  Matching  Content Models

Personalization by modeling users and content

Long-term

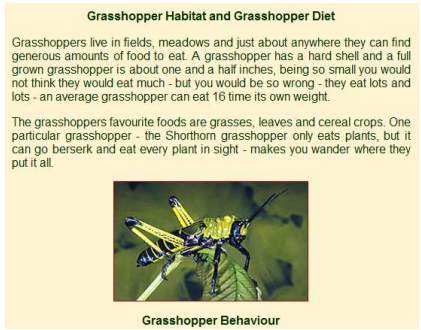
Short-term (this talk)

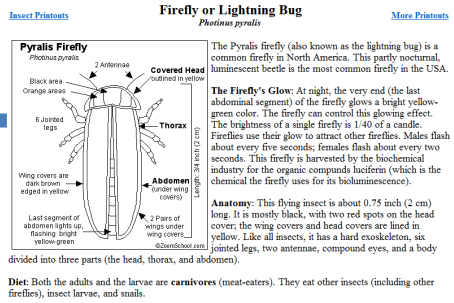
User and Intent

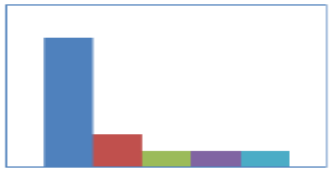


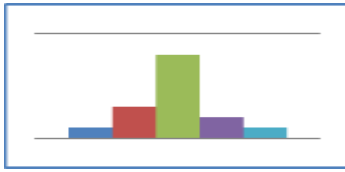
Desired reading level

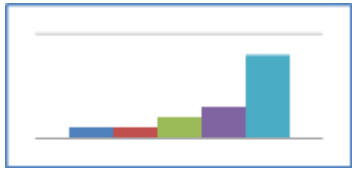
Re-ranker











Content reading level

How could a Web search engine personalize results by reading level?

1. Model a user's likely search intent:
   - Get explicit preferences or instructions from a user
   - Learn a user's interests and expertise over time
2. Extract reading-level and topical features:
   - Queries and Sessions: (Query text, results clicked, …)
   - User Profile (Explicit or Implicit from history)
   - Page reading level, Result snippet level
3. Use these features for personalized re-ranking

A simple session model combines the reading levels of previous satisfied clicks

Session reading level distribution

Typical features used for reading level personalization

- Content
  - Page reading level (*query-agnostic*)
  - Result snippet reading level (*query-dependent*)
- User: Session
  - Reading level averaged across previous satisfied clicks
  - Count of previous queries in session
- User: Query
  - Length in words, characters
  - Reading level prediction for raw text
- Interaction features
  - Snippet-Page, Query-Page, Query-Snippet
- Confidence features for many of the above

What types of queries are helped most by reading level personalization?

Query subset	Num. queries	% Total	Gain
Kids	15,796	4.4%	+1.0*
Science	23,059	6.8%	+4.3*
Sports	41,139	11.6%	+1.0*
Health	21,581	6.1%	0.0
All	545,255	100%	+1.2*

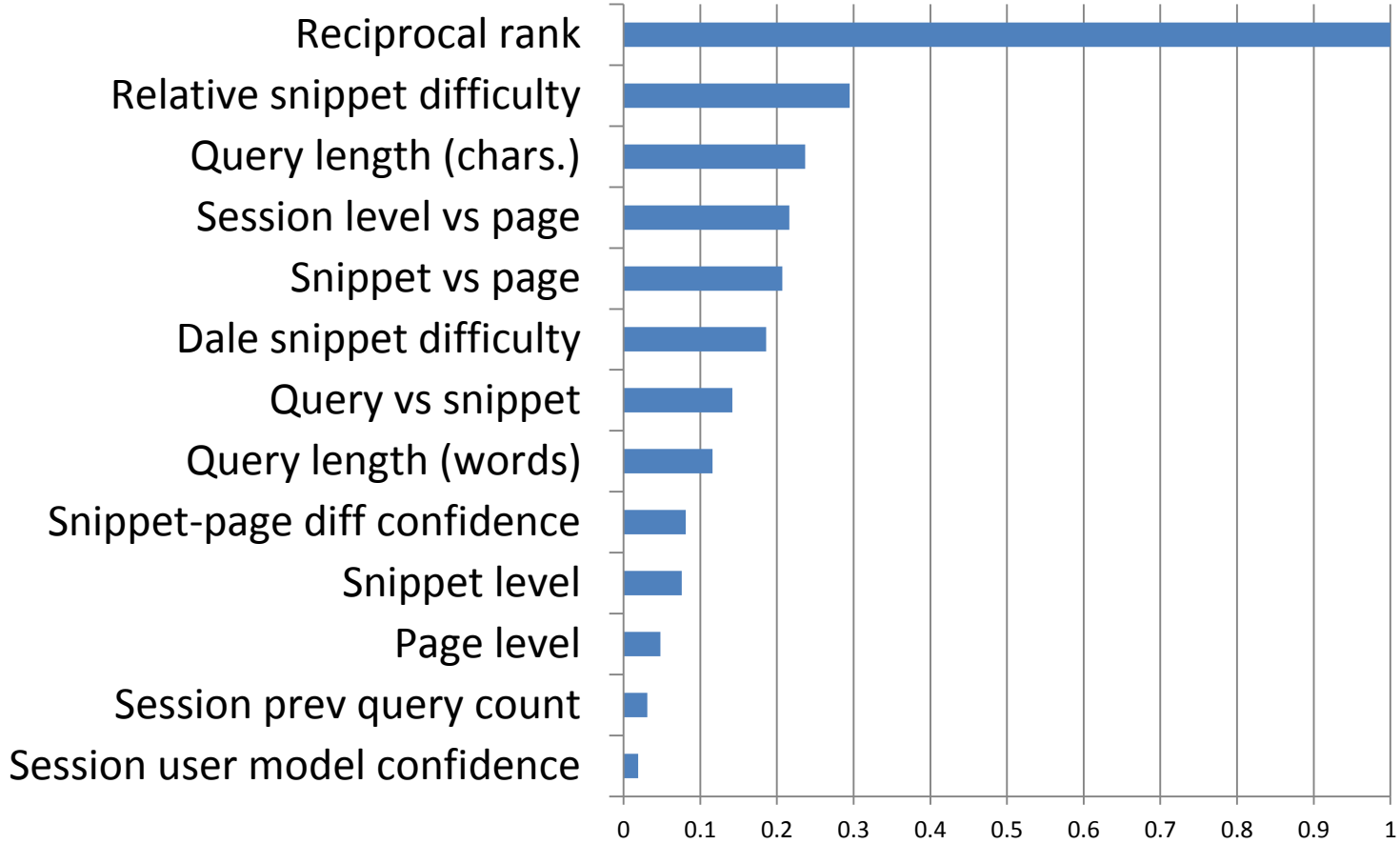
Point-Gain in Mean Reciprocal Rank of Last-SAT click

- Gain for all queries, and most query subsets (205, 623 sessions)
  - Size of gain varied with query subset
  - Science queries benefited most in our experiment
- Beating the default production baseline is very hard: Gain ≥ 1.0 is notable
- Net +1.6% of all queries improved at least one rank position in satisfied click
  - Large rank changes (> 5 positions) more than 70% likely to result in a win

What features were most important for reading level personalization?

- Session-based context
  - Results that match the reading level of previously clicked results in a user's session
- Good snippet-page match
  - The result snippet should faithfully represent the difficulty of the page
- Low relative snippet difficulty
  - Users prefer easiest snippet, all things being equal
- Query length in characters
  - Captures longer single terms: better than word count
- Using all features performed best

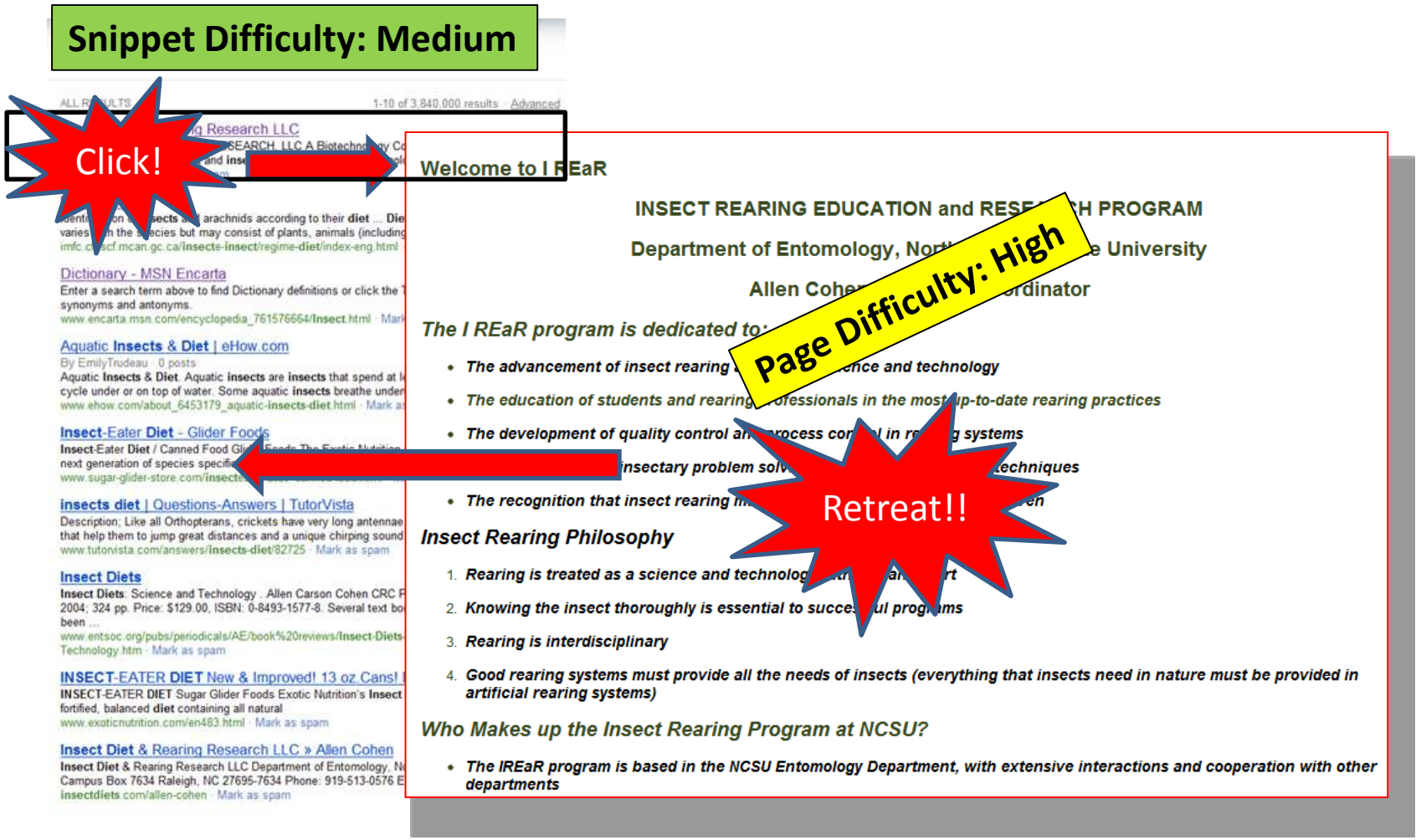
What features were most important for reading level personalization?



Average reduction in residual squared error over all trees and over all splits relative to the most informative feature.

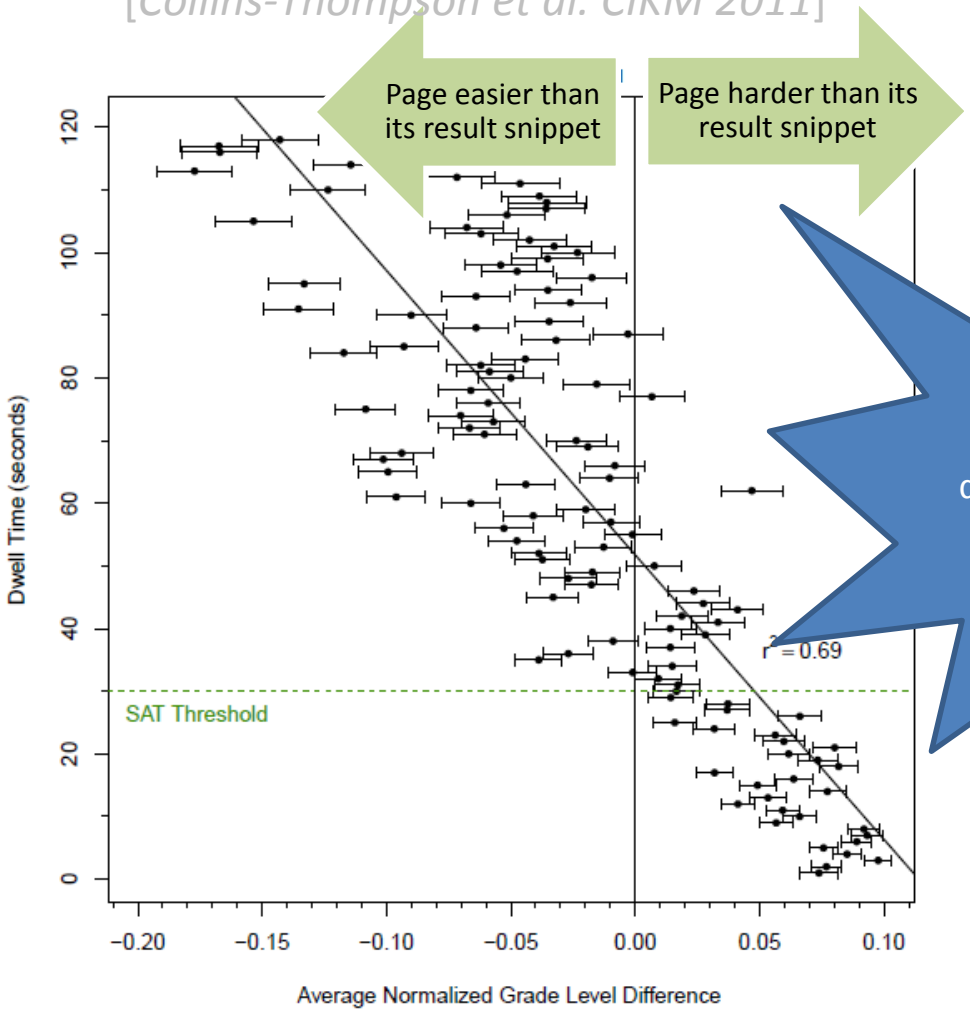
Application:
Improving snippet quality

Users can be misled by a mismatch between snippet readability and page readability



Users abandon pages faster when actual page is more difficult than the search result snippet suggested

[*Collins-Thompson et al. CIKM 2011*]



Future goal:
Expected snippet difficulty should match the underlying document difficulty

Application:

Modeling expertise on the Web using reading level + topic metadata

[*Kim, Collins-Thompson, Bennett, Dumais: WSDM 2012*]

Topic drift can occur when the specified reading level changes

Example: [quantum theory]

Quantum mechanics - Wikipedia, the free encyclopedia
History · Mathematical formulations · Mathematically ... · Interactions with ...
Quantum mechanics (QM - also known as **quantum** physics, or **quantum theory**) is a branch of physics dealing with physical phenomena where the action is on the order ...
en.wikipedia.org/wiki/**Quantum**_mechanics

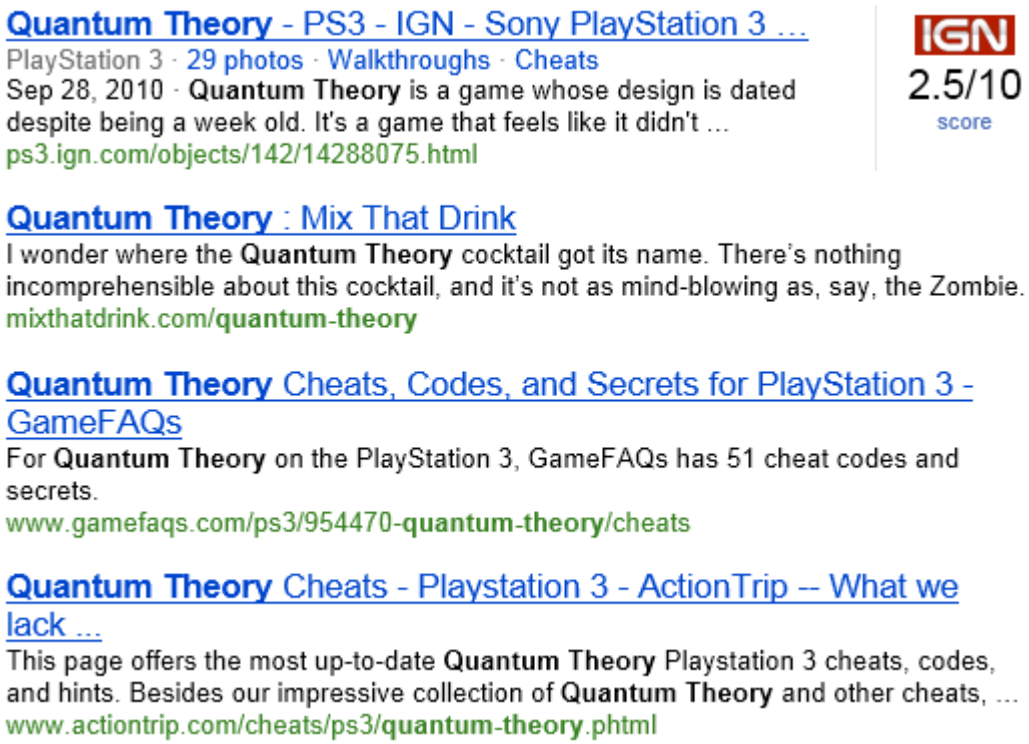
quantum theory: Definition from Answers.com
quantum theory n. A **theory** in physics based on the principle that matter and energy have the properties of both particles and waves, created to explain
www.answers.com/topic/**quantum-theory**

Quantum theory - Wikipedia, the free encyclopedia
Quantum theory may mean: In science: **Quantum** mechanics: a subset of **quantum** physics explaining the physical behaviours at atomic and sub-atomic levels Old **quantum** ...
en.wikipedia.org/wiki/**Quantum_theory**

Quantum Theory - thebigview.com - Pondering the Big Questions
Discovering the fundamental structure of matter. **Quantum theory** evolved as a new branch of theoretical physics during the first few decades of the 20th century in an ...
www.thebigview.com/spacetime/**quantumtheory**.html

Top 4 results

[quantum theory] + lower difficulty



Top 4 results

[quantum theory] + lower difficulty + science topic constraint

Quantum Theory
Quantum theory as a science is officially dead and has been replaced by multiple facets that include such things as **quantum** mechanics. These multi-faceted points ...
www.**quantumtheory**.org

Does **Quantum Theory** Explain Consciousness? : Discovery News
Just because consciousness is a mystery and **quantum theory** is mysterious, it doesn't mean they're connected.
news.discovery.com/space/does-**quantum-theory**-explain-consciousness...

Quantum Theory | PlanetSEED
The Expanding Universe **Quantum Theory** Einstein's Big Mistake? Another big problem goes right back to the way Einstein guessed his equations in 1917.
https://www.planetseed.com/.../the-expanding-universe/**Quantum-Theory**

Einstein's Intuition : **Quantum** Space **Theory**
Einstein's Intuition : **Quantum** Space **Theory**: ... Questions and answers: I'd like to dedicate this page to questions that anyone out there might have regarding

Top 4 results

[cinderella] + higher difficulty

Cinderella : Cinderella
Cinderella is a Java based interactive geometry tool. The only available tool that gives correct solutions to typical geometrical problems.
www.**cinderella**.de

Cinderella Software
If you only need to browse and/or print SDL files, then download our free viewing tool. **Cinderella** SDL is a visual modeling tool for developing embedded software ...
www.**cinderella**.dk/index.htm

Cinderella - School of Ballroom Dancing
About Us: Home | Contact Us : W elcome to the **Cinderella** School of Ballroom Dancing. Ballroom dancing is as romantic as it is enjoyable. For years the world's ...
cinderelladanceschool.com/index.htm

Interactives . Elements of a Story . **Cinderella**
About this Interactive | Tips for Adults | Elements of a Story Site Map
www.learner.org/interactives/story/**cinderella**.html

Cinderella
I bought **Cinderella**, and it is running in German only. I have a Mac. **Cinderella** does not run on my Computer, although I know I have Java 2 installed on it [usually ...
cinderella.de/tiki-view_faq.php?faqId=1

Top 4 results

[bambi]

Bambi - Wikipedia, the free encyclopedia
Plot · Cast · Production · Release · Reception · Legacy
Bambi is a 1942 American animated film directed by David Hand (supervising a team of sequence directors), produced by Walt Disney and based on the book **Bambi**, A ...
en.wikipedia.org/wiki/**Bambi**

Images of **bambi**
See also: Bambi 2 Vhs · Disney Images Of Bambi · Bambi And Feline

Bambi (1942) - IMDb
Animation/Drama/Family · 70 min
Director: James Algar, Samuel Armstrong. . Actors: Hardie Albright: Adolescent **Bambi** · Stan Alexander: Young Flower · Bobette ...
www.imdb.com/title/tt0034492

IMDb
7.5/10
39,633 ratings

Top 3 results

[bambi] + higher difficulty

The SETI-Capable BAMBI Radio Telescope

By Bob Lash and Mike Fremont

Introduction

A number of efforts are underway in the Search for Extraterrestrial Intelligence (SETI). We have been deeply interested in the search for some time, and have concluded that amateurs can in fact construct affordable systems with sensitivities comparable to professional all-sky search strategies even with antennas of limited aperture. We have also concluded that we can achieve a reasonably respectable frequency coverage of a search spectrum as well. We hope this project will encourage other amateurs to join in the search. Project BAMBI is divided into two phases:

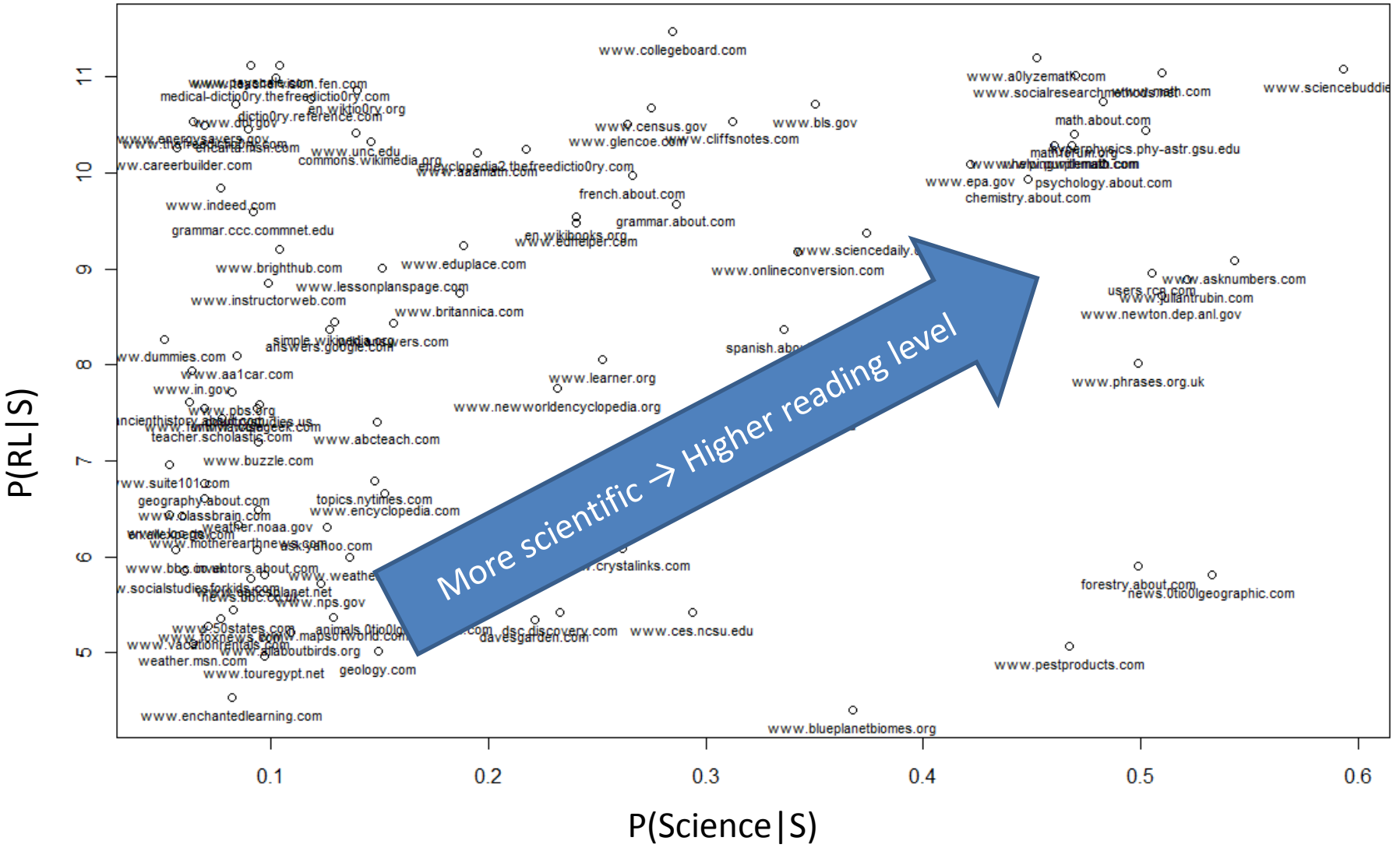
Phase I: Standard Amateur Radio Astronomy:

We have initially operated BAMBI as a total power receiver for several

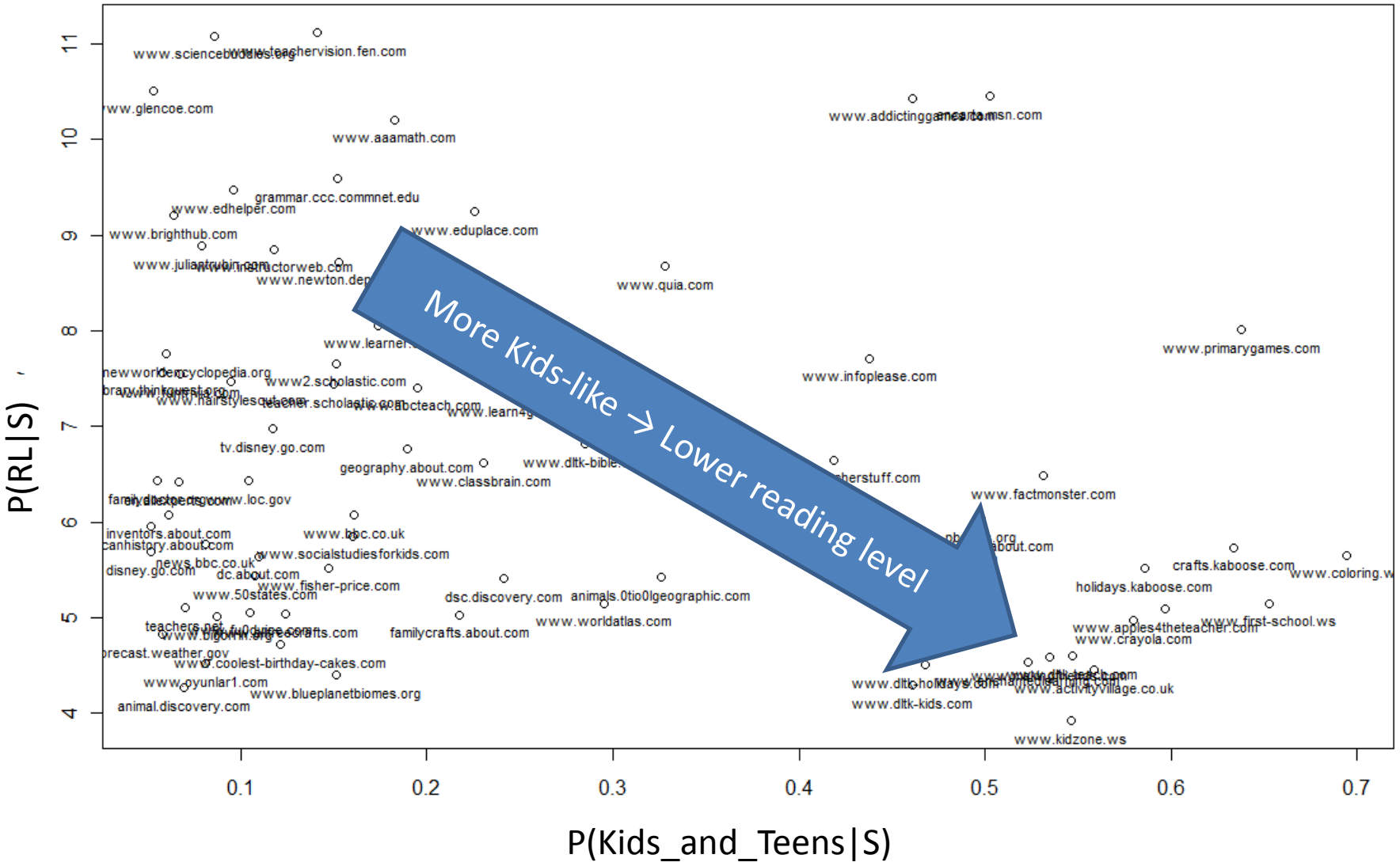
P(RL|T) for Top ODP Topic Categories

Top Category	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	E(RL)
Home	0.00	0.00	0.02	0.30	0.45	0.08	0.03	0.01	0.01	0.01	0.07	0.02	5.49
Shopping	0.00	0.00	0.01	0.16	0.32	0.23	0.10	0.04	0.02	0.03	0.07	0.02	6.14
Recreation	0.00	0.00	0.01	0.11	0.43	0.19	0.09	0.03	0.01	0.02	0.08	0.02	6.15
Sports	0.00	0.00	0.00	0.09	0.48	0.12	0.12	0.04	0.02	0.02	0.08	0.02	6.19
News	0.00	0.00	0.00	0.06	0.42	0.18	0.17	0.03	0.01	0.01	0.08	0.03	6.36
Arts	0.00	0.00	0.01	0.10	0.37	0.15	0.14	0.06	0.01	0.02	0.09	0.04	6.48
Kids_and_Teens	0.00	0.00	0.02	0.19	0.32	0.13	0.09	0.03	0.01	0.03	0.11	0.07	6.54
Adult	0.00	0.00	0.00	0.07	0.28	0.26	0.15	0.06	0.01	0.01	0.09	0.06	6.73
Games	0.00	0.00	0.01	0.13	0.29	0.13	0.11	0.04	0.02	0.03	0.19	0.05	7.09
Society	0.00	0.00	0.00	0.07	0.31	0.14	0.11	0.06	0.02	0.03	0.16	0.08	7.27
Business	0.00	0.00	0.01	0.07	0.23	0.18	0.09	0.03	0.02	0.04	0.22	0.11	7.74
Science	0.00	0.00	0.00	0.06	0.23	0.09	0.07	0.02	0.01	0.07	0.27	0.17	8.46
Reference	0.00	0.00	0.00	0.03	0.17	0.10	0.16	0.04	0.02	0.03	0.23	0.21	8.61
Health	0.00	0.00	0.00	0.03	0.16	0.07	0.13	0.04	0.03	0.11	0.30	0.13	8.79
Computers	0.00	0.00	0.00	0.04	0.10	0.07	0.05	0.02	0.01	0.04	0.43	0.23	9.62

P(RL|S) against P(Science|S)

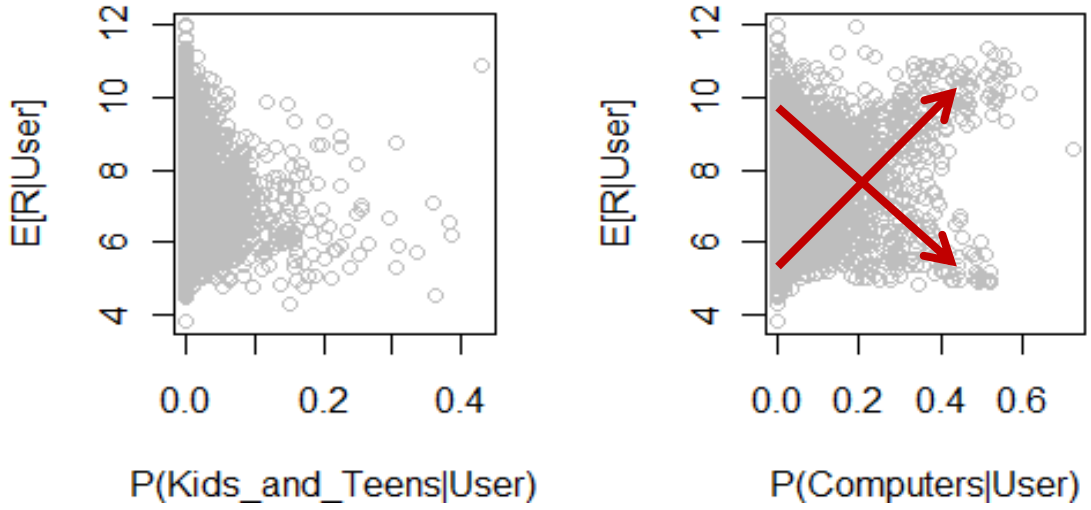


P(RL|S) against P(Kids_and_Teens|S)



User Reading Level against P(Topic)

- Results suggest that there are both expert (high RL) and novice (low RL) users for computer topics



Using reading level and topic <u>together</u> to model user and site expertise

Four features that aggregate metadata over pages:

Reading level:

1. Expected reading level $E(R)$ over site/user pages
2. Entropy $H(R)$ of reading level over site/user pages

Topic:

3. Top-K ODP category predictions over site/user pages
4. Entropy $H(T)$ of ODP category distribution for site/user pages

Sites with low topic entropy (focused) tend to be expert-oriented

Sites with focused topical content: Low Entropy, H(T|S) < 1

Website	H(T|S)	T1	P1	T2	P2	T3	P3
www.prosportsdaily.com	0.83	Sports	0.74	Sports/Football	0.26		
www.organize.com	0.91	Shopping	0.67	Shop/Home&Garden	0.33		
www.trulia.com	0.92	Business	0.78	Society	0.18	Bus./Construction	0.04
www.fandango.com	0.95	Arts	0.63	Arts/Movies	0.36		
www.hobbytron.com	0.96	Recreation	0.62	Shopping	0.38		

Sites with high topic entropy (breadth) tend to be for general audiences

Sites with focused topical content: Low Entropy, H(T|S) < 1

Website	H(T|S)	T1	P1	T2	P2	T3	P3
www.prosportsdaily.com	0.83	Sports	0.74	Sports/Football	0.26		
www.organize.com	0.91	Shopping	0.67	Shop/Home&Garden	0.33		
www.trulia.com	0.92	Business	0.78	Society	0.18	Bus./Construction	0.04
www.fandango.com	0.95	Arts	0.63	Arts/Movies	0.36		
www.hobbytron.com	0.96	Recreation	0.62	Shopping	0.38		

Sites with very broad topical content: High Entropy : H(T|S) > 4

Website	H(T|S)	T1	P1	T2	P2	T3	P3
ezinearticles.com	4.27	Business	0.12	Health	0.09	Home	0.08
www.dummies.com	4.28	Computers	0.17	Computers/HW	0.09	Business	0.08
en.allexperts.com	4.38	Recreation	0.12	Home	0.09	Recreation/Pets	0.07
phoenix.about.com	4.38	Recreation	0.12	Society	0.09	Arts	0.07
www.wisegeek.com	4.40	Health	0.12	Business	0.10	Science	0.09

Reading level entropy measures breadth of a site's content difficulty

Sites with focused reading level: Low Entropy, H(RL|S) < 1

Website	H(RL|S)	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	Count	E(RL|S)
www.pumpkinpatchesandmore.org	0.99	0	0	0.7	0.2	0	0	0	0	0	0	0	0	35	3.3
busycooks.about.com	0.9	0	0	0	0.8	0.1	0	0	0	0	0	0	0	45	4.12
www.pickyourown.org	0.93	0	0	0	0.8	0.2	0	0	0	0	0	0	0	38	4.14
www.ssa.gov	0.91	0	0	0	0	0	0	0	0	0	0	0.1	0.8	59	11.52
h10025.www1.hp.com	0.78	0	0	0	0	0	0	0	0	0	0	0.2	0.8	55	11.77
www.socialsecurity.gov	0.53	0	0	0	0	0	0	0	0	0	0	0.1	0.9	29	11.87

Sites with broad range of reading level: High Entropy, H(RL|S) > 2

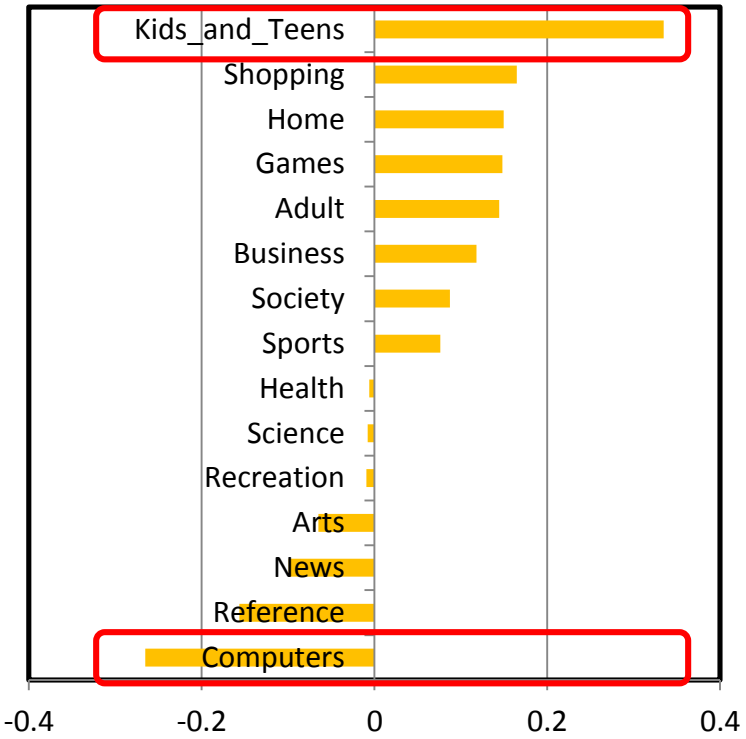
Website	H(RL|S)	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	Count	E(RL|S)
www.dltk-kids.com	2.02	0	0	0.2	0.5	0.2	0.1	0	0	0	0	0	0	39	4.4
www.dltk-teach.com	2.1	0	0	0.2	0.4	0.2	0.2	0	0	0	0	0	0	26	4.47
www.dltk-holidays.com	2.07	0	0	0.2	0.5	0.1	0	0.1	0	0	0	0	0	31	4.65
psychology.about.com	2.32	0	0	0	0	0	0	0.1	0	0	0.2	0.3	0.4	59	10.46
compnetworking.about.com	2.07	0	0	0	0	0	0	0.1	0	0	0.1	0.4	0.4	68	10.58
pcsupport.about.com	2.02	0	0	0	0	0	0	0	0	0	0.1	0.4	0.3	39	10.68

Site Reading Level vs. Visitor Diversity

- Expected reading level of site is uncorrelated with visitor diversity

Website Reading Level	Visitor Profile Diversity		
	$Div_R(U|s)$	$Div_T(U|s)$	$Div_{RT}(U|s)$
E[R|s]	0.052	0.081	0.095

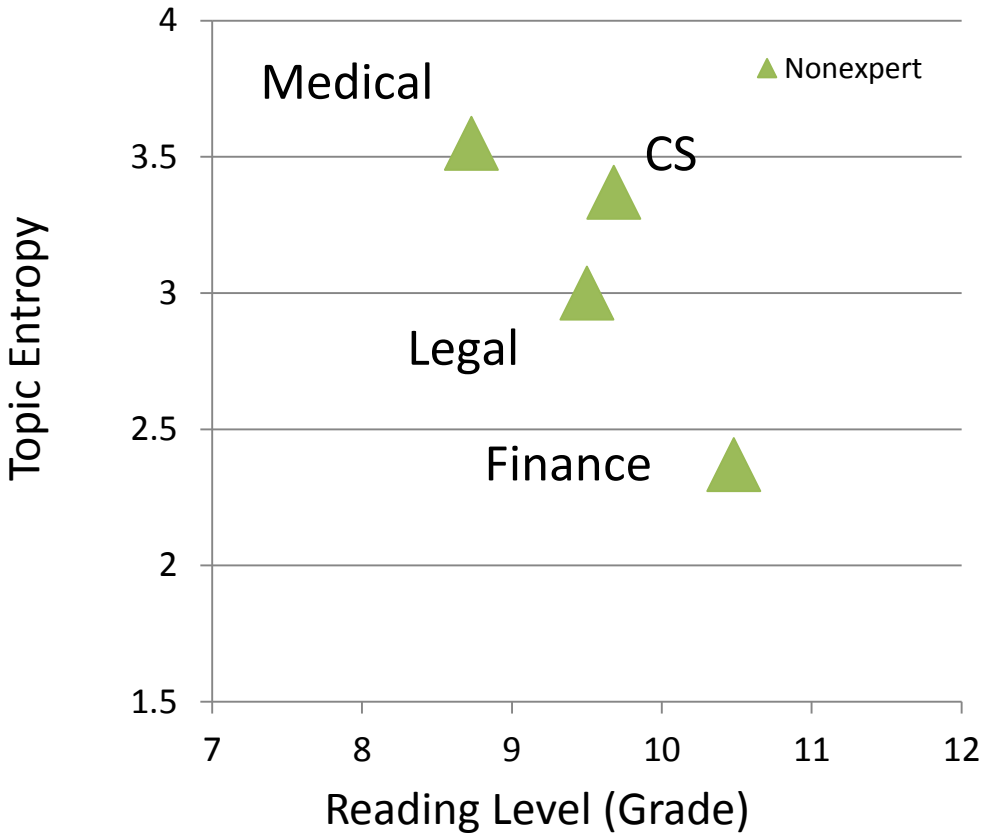
- …But a breakdown of sites by topic reveals stronger relationships
  - Computer sites with high reading level attract focused visitors
  - Kids sites with high reading level attract diverse visitors



(RL, Div) Correlation by Site Topic

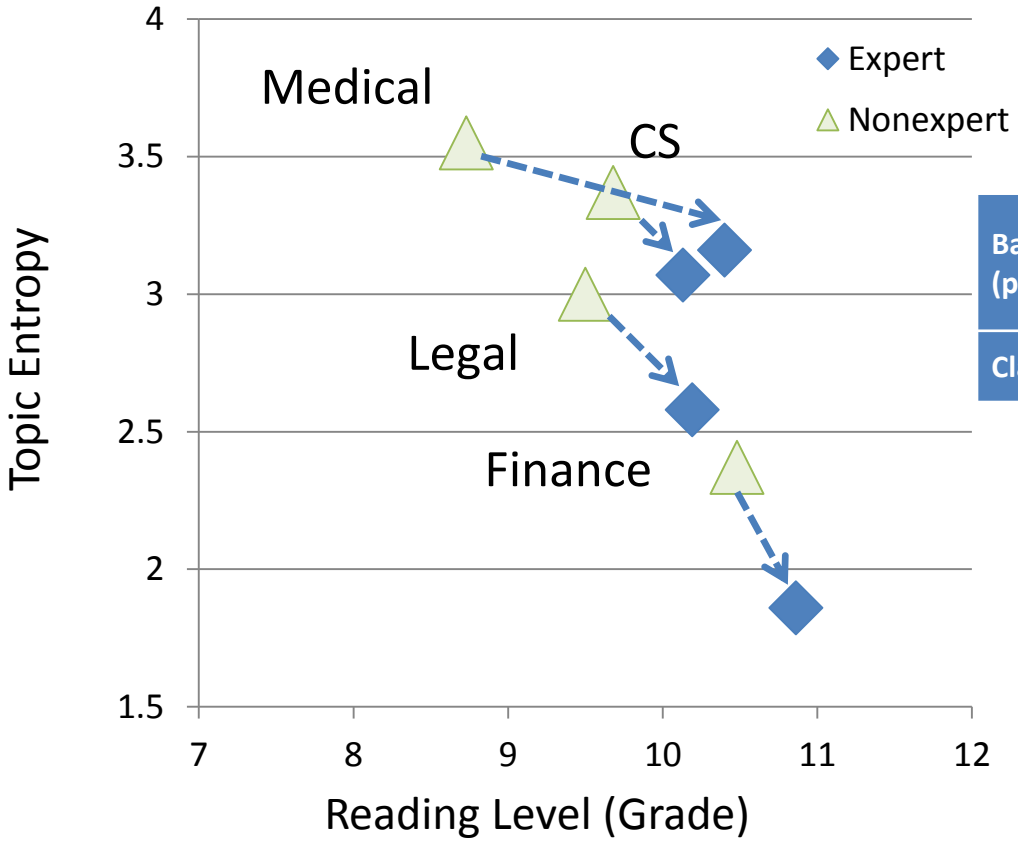
Reading level and topic entropy features can help separate expert from non-expert websites

[Kim, Collins-Thompson, Bennett, Dumais. WSDM 2012]



Reading level and topic entropy features can help separate expert from non-expert websites

[Kim, Collins-Thompson, Bennett, Dumais. WSDM 2012]



Baseline (predict most likely class)	65.8%
Classifier accuracy	82.2%

Application: Searcher motivation

Readability metadata may also help predict when searchers are highly motivated

- Sites that are popular but also have large difference from average reading level

Website	Type of site
socialsecurity.gov	Government retirement/disability
collegeboard.com	Entrance exam preparation, college application help
softwarepatch.com	Find software patches
fileinfo.com	Find programs to open file types
msdn.microsoft.com	Technical reference

‘Stretch’ tasks: what are people searching for when they *deviate* from their typical reading level profile?

Capturing stretch behaviors:

- Estimate a user’s typical reading level profile over time, from historical search data
- Collect search sessions where

  $$E[R|Session] - E[R|User] > 4 \text{ grade levels}$$

- Build language models from titles of clicked pages
- Compare word probability in clicked vs. all titles

‘Stretch’ tasks: what are people searching for when they *deviate* from their typical reading level profile?

Highest association with stretch reading	
Title word	Log ratio
tests	2.22
test	1.99
sample	1.94
digital	1.88
options	1.87
aid	1.87
effects	1.84
education	1.77
forms	1.76
plan	1.74
pay	1.71
medical	1.69
learning	1.62

Medical tests
College entrance

Financial aid

Gov’t forms
Job search

[Kim et al, WSDM 2012] Based on 2-month user profiles from Bing search log data

‘Stretch’ tasks: what are people searching for when they *deviate* from their typical reading level profile?

Highest association with stretch reading		Lowest association with stretch reading	
Title word	Log ratio	Title word	Log ratio
tests	2.22	best	-0.42
test	1.99	football	-0.45
sample	1.94	store	-0.46
digital	1.88	great	-0.47
options	1.87	items	-0.52
aid	1.87	new	-0.53
effects	1.84	sale	-0.61
education	1.77	games	-0.65
forms	1.76	sports	-0.78
plan	1.74	food	-0.81
pay	1.71	news	-0.82
medical	1.69	music	-1.02
learning	1.62	all	-1.35

Medical tests
College entrance

Financial aid

Gov’t forms

Shopping!
Exploration
Leisure

[Kim et al, WSDM 2012] Based on 2-month user profiles from Bing search log data

‘Stretch’ tasks: what are people searching for when they *deviate* from their typical reading level profile?

Medical tests
College entrance

Financial aid

Gov’t forms

Shopping!
Exploration
Leisure

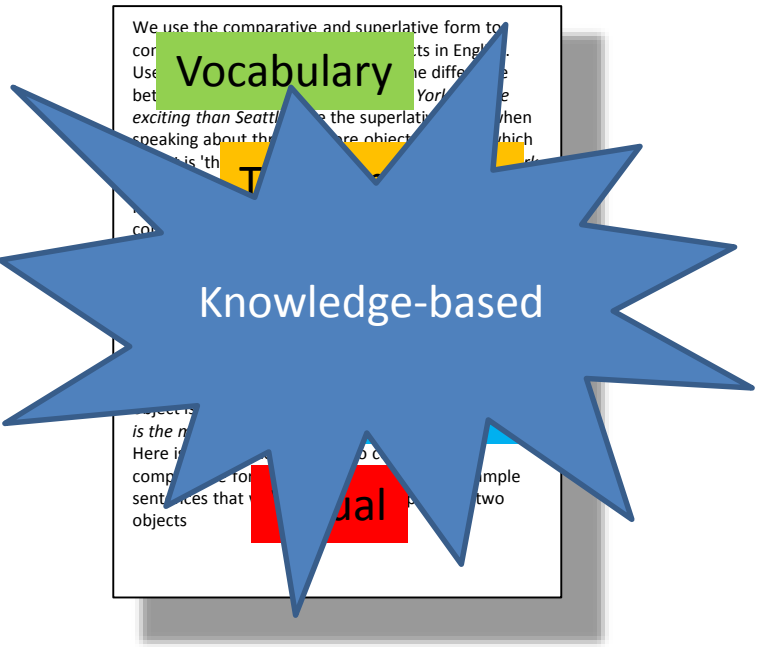
Highest association with stretch reading		Lowest association with stretch reading	
Title word	Log ratio	Title word	Log ratio
tests	2.1	[illegible]	[illegible]
test			
sample			
[illegible]			
option			
aid			
effect			
[illegible]			
forms		[illegible]	[illegible]
plan	1.7	[illegible]ood	-0.81
pay	1.7	news	-0.82
medical	1.69	music	-1.02
learning	1.62	all	-1.35

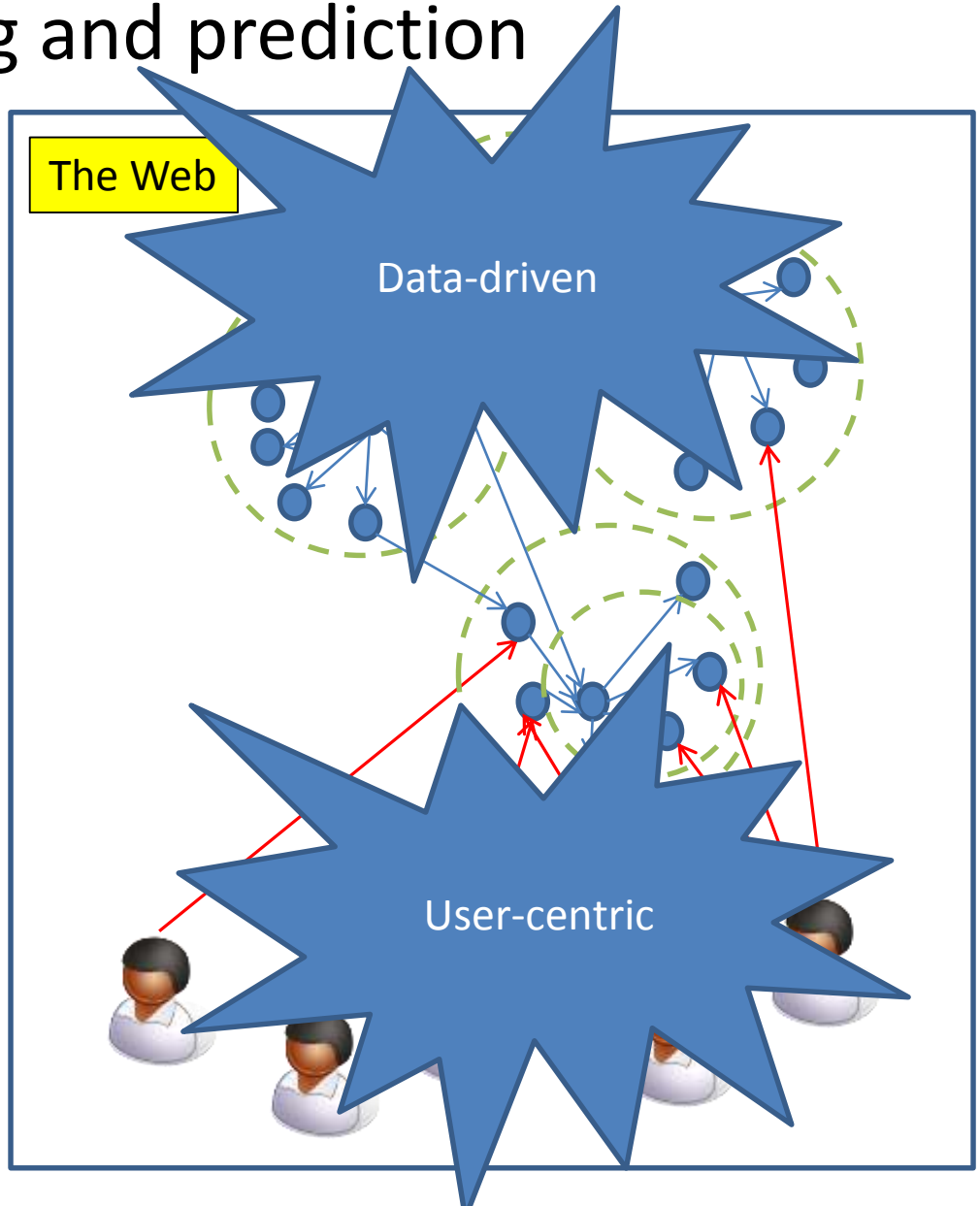
Future work:

1. Identify & predict stretch tasks
2. Decide how and when to provide support
3. Determine helpful background or alternatives

[Kim et al, WSDM 2012] Based on 2-month user profiles from Bing search log data

Three key innovation directions for readability modeling and prediction





Some key challenges and opportunities for readability research

Basic Advancement of Knowledge

• Deep content understanding - Identifying gaps and assumptions - Concepts and their dependencies • Deep user understanding - Your expertise & changes over time - Learning plans tailored for you - Cognitive models of learning	• Data-driven, personalized readability measures • Adapting content to users - Enrich, augment, rewrite • Adapting users to content • Influencing search presentation and interaction
• Analyzing movie scripts with Keanu Reeves dialogue 	• Web-scale speed and reliability • Exploiting new content forms - Blogs, wiki structure & edits • Adapting to different tasks and populations • Human computation/crowdsource • Predicting quality/authority

Relevance for applications

Next practical steps

- Working on adding rich reading-level features to ClueWeb09 and ClueWeb12
- Applications to learning analytics
  - Text mining of Univ of Michigan student content
- Crowdsourced expertise/difficulty annotation

Thanks! Questions?

For more information:

E-mail: kevynct@umich.edu

Web site:

`http://www.umich.edu/~kevynct`