

# 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



# 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*)

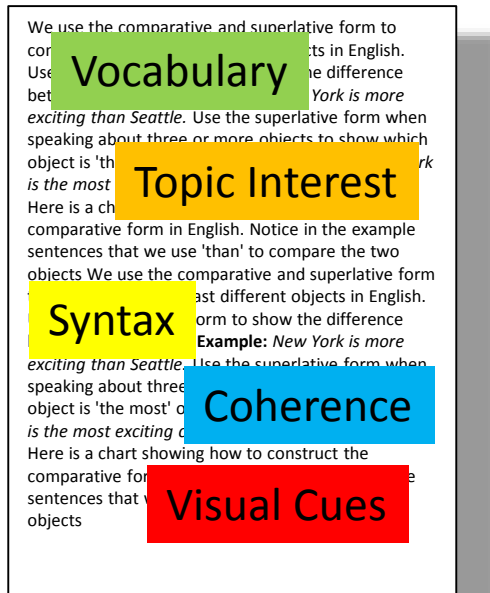
Guy Lebanon (*Amazon*)

David Sontag (*NYU*)

# Bringing together readability and the Web

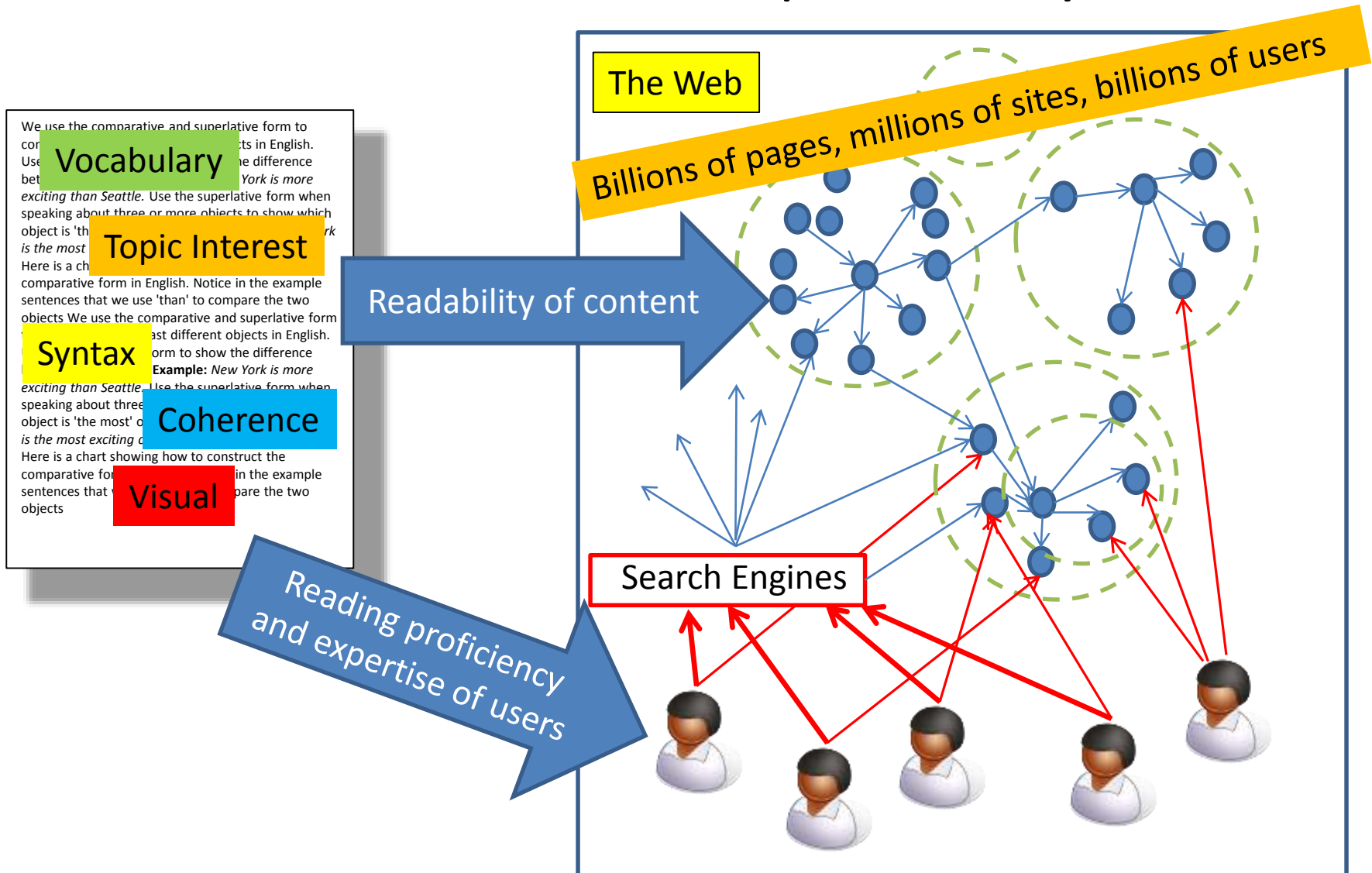
## ... sometimes in unexpected ways

### Text Readability Modeling and Prediction

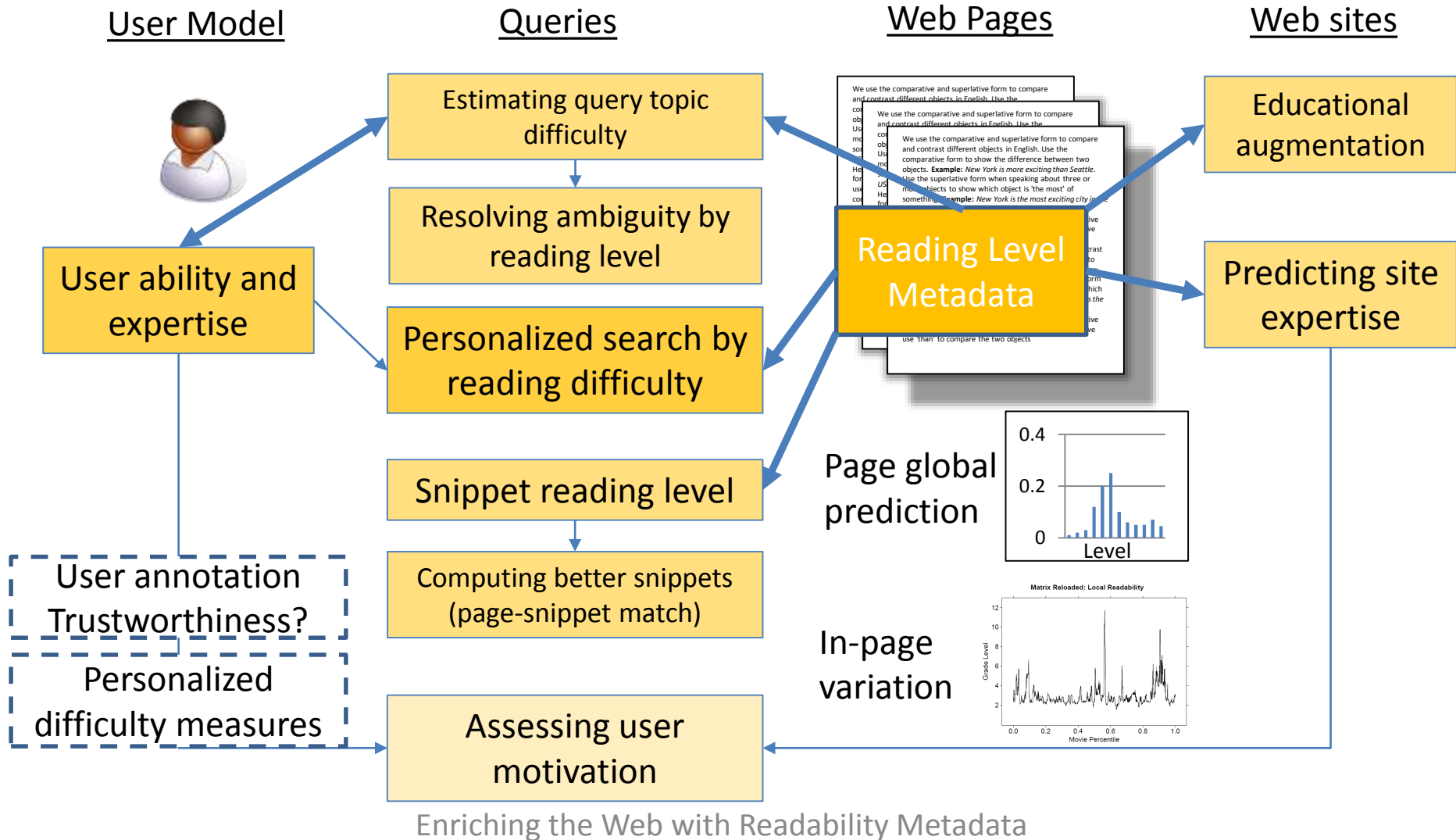


Reading level prediction  
Topic prediction

# 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 times 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 wonder where they put it all.



## Grasshopper Behaviour

Query [insect diet]: Lower difficulty

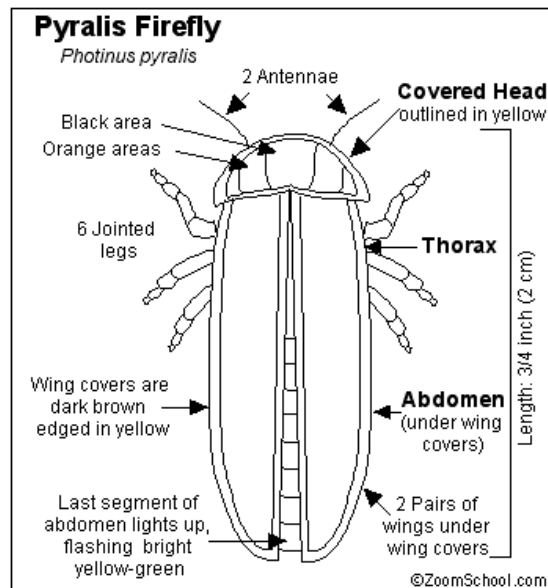
# Medium difficulty [insect diet]

[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 compounds 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

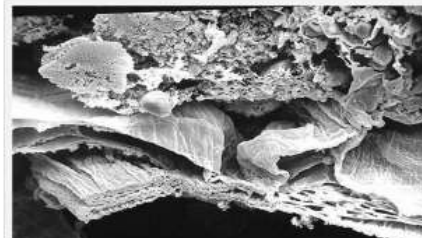
### REARING RESEARCH ON DIET DEVELOPMENT AND ESTABLISHING NEW AND IMPROVED REARING SYSTEMS

- 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]

insect diet 

ALL RESULTS 1-10 of 3,840,000 results - [Advanced](#)

**[Insect Diet & Rearing Research LLC](#)**  
INSECT DIET and REARING RESEARCH, LLC A Biotechnology Company dedicated to: The advancement of **insect diet** and **insect** rearing science and technology through education, ...  
[insectdiets.com](#) · [Mark as spam](#)

**[Insect diet](#)**  
Identification of **insects** and arachnids according to their **diet** ... **Diet**. The **diet** of arthropods varies with the species but may consist of plants, animals (including other ...  
[imfc.cfl.scf.mcan.gc.ca/insecte-insect/regime-diet/index-eng.html](#) · [Mark as spam](#)

**[Dictionary - MSN Encarta](#)**  
Enter a search term above to find Dictionary definitions or click the Thesaurus tab to find synonyms and antonyms.  
[www.encarta.msn.com/encyclopedia\\_761576664/Insect.html](#) · [Mark as spam](#)

**[Aquatic Insects & Diet | eHow.com](#)**  
By EmilyTrudeau · 0 posts  
**Aquatic Insects & Diet** Aquatic **insects** are **insects** that spend at least a portion of their life cycle under or on top of water. Some aquatic **insects** breathe underwater with gills ...  
[www.ehow.com/about\\_6453179\\_aquatic-insects-diet.html](#) · [Mark as spam](#)

**[Insect-Eater Diet - Glider Foods](#)**  
**Insect-Eater Diet / Canned Food Glider Foods** The Exotic Nutrition company has pioneered the next generation of species specific **diets** for exotics. The  
[www.sugar-glider-store.com/insecteater-diet-canned-food.html](#) · [Mark as spam](#)

**[insects diet | Questions-Answers | TutorVista](#)**  
Description: Like all Orthopterans, crickets have very long antennae and long, strong back legs that help them to jump great distances and a unique chirping sound that they ...  
[www.tutortvta.com/answers/insects-diet/82725](#) · [Mark as spam](#)

**[Insect Diets](#)**  
**Insect Diets**: Science and Technology . Allen Carson Cohen CRC Press LLC Boca Raton, FL 2004; 324 pp. Price: \$129.00, ISBN: 0-8493-1577-8. Several text books and monographs have been ...  
[www.entsoc.org/pubs/periodicals/AE/book%20reviews/Insect-Diets-Science-and-Technology.htm](#) · [Mark as spam](#)

**[INSECT-EATER DIET New & Improved! 13 oz Cans! Moist & Easier to ...](#)**  
**INSECT-EATER DIET** Sugar Glider Foods Exotic Nutrition's **Insect Eater Diet** TM (13 oz can) is a fortified, balanced **diet** containing all natural  
[www.exoticonnutrition.com/en483.html](#) · [Mark as spam](#)

**[Insect Diet & Rearing Research LLC » Allen Cohen](#)**  
**Insect Diet & Rearing Research LLC** Department of Entomology, North Carolina State University Campus Box 7634 Raleigh, NC 27695-7634 Phone: 919-513-0576 Email: [idr@insectdiets.com](mailto:idr@insectdiets.com)  
[insectdiets.com/allen-cohen](#) · [Mark as spam](#)

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

insect diet

ALL RESULTS 1-10 of 3,840,000 results - Advanced

**[Insect Diet & Rearing Research LLC](#)**  
INSECT DIET and REARING RESEARCH, LLC A Biotechnology Company dedicated to: The advancement of **insect diet** and **insect** rearing science and technology through education, ...  
[insectdiets.com](#) · Mark as spam

**[Insect diet](#)**  
Identification of **insects** and arachnids according to their **diet** ... **Diet**. The **diet** of arthropods varies with the species but may consist of plants, animals (including other ...  
[imfc.cfl.scf.mcan.gc.ca/insecte-insect/regime-diet/index-eng.html](#) · Mark as spam

**[Dictionary - MSN Encarta](#)**  
Enter a search term above to find Dictionary definitions or click the Thesaurus tab to find synonyms and antonyms.  
[www.encarta.msn.com/encyclopedia\\_761576664/Insect.html](#) · Mark as spam

**[Aquatic Insects & Diet | eHow.com](#)**  
By EmilyTrudeau · 0 posts  
Aquatic **Insects & Diet** Aquatic **insects** are **insects** that spend at least a portion of their life cycle under or on top of water. Some aquatic **insects** breathe underwater with gills ...  
[www.ehow.com/about\\_6453179\\_aquatic-insects-diet.html](#) · Mark as spam

**[Insect-Eater Diet - Glider Foods](#)**  
**Insect-Eater Diet** / Canned Food Glider Foods The Exotic Nutrition company has pioneered the next generation of species specific **diets** for exotics. The  
[www.sugar-glider-store.com/insecteater-diet-canned-food.html](#) · Mark as spam

**[insects diet | Questions-Answers | TutorVista](#)**  
Description: Like all Orthopterans, crickets have very long antennae and long, strong back legs that help them to jump great distances and a unique chirping sound that they ...  
[www.tutortvta.com/answers/insects-diet/82725](#) · Mark as spam

**[Insect Diets](#)**  
**Insect Diets**: Science and Technology . Allen Carson Cohen CRC Press LLC Boca Raton, FL 2004; 324 pp. Price: \$129.00, ISBN: 0-8493-1577-8. Several text books and monographs have been ...  
[www.entsoc.org/pubs/periodicals/AE/book%20reviews/Insect-Diets-Science-and-Technology.htm](#) · Mark as spam

**[INSECT-EATER DIET New & Improved! 13 oz Cans! Moist & Easier to ...](#)**  
**INSECT-EATER DIET** Sugar Glider Foods Exotic Nutrition's **Insect Eater Diet** TM (13 oz can) is a fortified, balanced **diet** containing all natural  
[www.exoticnutrition.com/en483.html](#) · Mark as spam

**[Insect Diet & Rearing Research LLC » Allen Cohen](#)**  
**Insect Diet & Rearing Research LLC** Department of Entomology, North Carolina State University Campus Box 7634 Raleigh, NC 27695-7634 Phone: 919-513-0576 Email: [idr@insectdiets.com](#)  
[insectdiets.com/allen-cohen](#) · Mark as spam

X Technical

X Technical

X Relevance



X Relevance

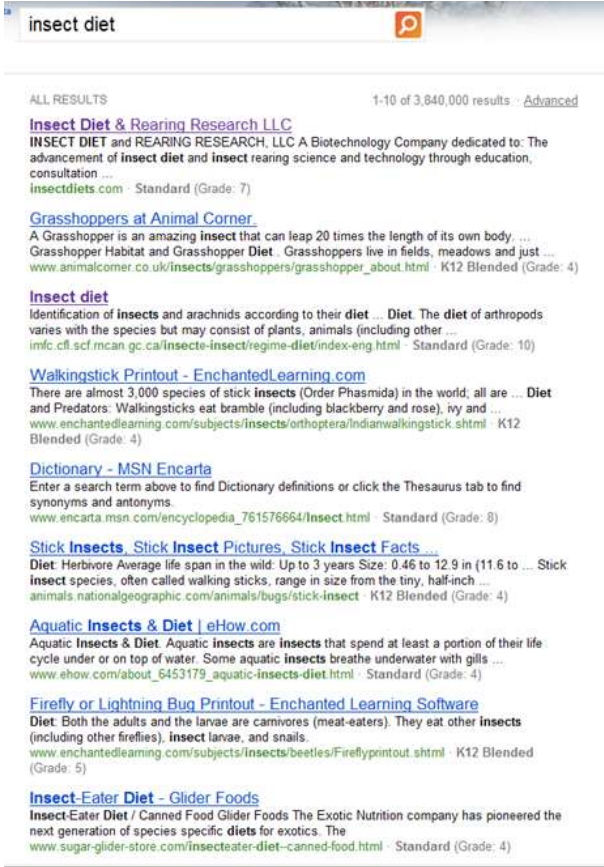


X Technical

X Relevance

X Technical

# Blending in lower difficulty results would improve relevance for this user



The image shows a search engine results page for the query "insect diet". The page displays several search results, each with a title, a brief description, and a URL. To the left of the results, there are five blue arrows pointing to specific entries. To the right, there are red "X" marks and smiley face icons indicating the relevance and technical difficulty of each result.

Result	Relevance	Technical
<a href="#">Insect Diet &amp; Rearing Research LLC</a> INSECT DIET and REARING RESEARCH, LLC A Biotechnology Company dedicated to: The advancement of <b>insect diet</b> and <b>insect</b> rearing science and technology through education, consultation ... <a href="#">insectdiets.com</a> · Standard (Grade: 7)	X Technical	😊
<a href="#">Grasshoppers at Animal Corner</a> A Grasshopper is an amazing <b>insect</b> that can leap 20 times the length of its own body. ... Grasshopper Habitat and Grasshopper <b>Diet</b> . Grasshoppers live in fields, meadows and just ... <a href="#">www.animalcorner.co.uk/insects/grasshoppers/grasshopper_about.html</a> · K12 Blended (Grade: 4)	X Technical	😊
<a href="#">Insect diet</a> Identification of <b>insects</b> and arachnids according to their <b>diet</b> ... <b>Diet</b> . The <b>diet</b> of arthropods varies with the species but may consist of plants, animals (including other ... <a href="#">imfc.cfl.scl.mcan.gc.ca/Insecte-Insect/regime-diet/index-eng.html</a> · Standard (Grade: 10)	X Relevance	😊
<a href="#">Walkingstick Printout - EnchantedLearning.com</a> There are almost 3,000 species of stick <b>insects</b> (Order Phasmoda) in the world, all are ... <b>Diet</b> and Predators: Walkingsticks eat bramble (including blackberry and rose), ivy and ... <a href="#">www.enchantedlearning.com/subjects/insects/orthoptera/indianwalkingstick.shtml</a> · K12 Blended (Grade: 4)	😊	😊
<a href="#">Dictionary - MSN Encarta</a> Enter a search term above to find Dictionary definitions or click the Thesaurus tab to find synonyms and antonyms. <a href="#">www.encarta.msn.com/encyclopedia_761576664/Insect.html</a> · Standard (Grade: 8)	X Relevance	😊
<a href="#">Stick Insects, Stick Insect Pictures, Stick Insect Facts ...</a> <b>Diet</b> : Herbivore Average life span in the wild: Up to 3 years Size: 0.46 to 12.9 in (11.6 to ... Stick <b>insect</b> species, often called walking sticks, range in size from the tiny, half-inch ... <a href="#">animals.nationalgeographic.com/animals/bugs/stick-insect</a> · K12 Blended (Grade: 4)	😊	😊
<a href="#">Aquatic Insects &amp; Diet   eHow.com</a> Aquatic <b>Insects &amp; Diet</b> . Aquatic <b>insects</b> are <b>insects</b> that spend at least a portion of their life cycle under or on top of water. Some aquatic <b>insects</b> breathe underwater with gills ... <a href="#">www.ehow.com/about_6453179_aquatic-insects-diet.html</a> · Standard (Grade: 4)	😊	😊
<a href="#">Firefly or Lightning Bug Printout - Enchanted Learning Software</a> <b>Diet</b> . Both the adults and the larvae are carnivores (meat-eaters). They eat other <b>insects</b> (including other fireflies), <b>insect</b> larvae, and snails. <a href="#">www.enchantedlearning.com/subjects/insects/beetles/Fireflyprintout.shtml</a> · K12 Blended (Grade: 5)	😊	😊
<a href="#">Insect-Eater Diet - Glider Foods</a> <b>Insect-Eater Diet</b> / Canned Food Glider Foods The Exotic Nutrition company has pioneered the next generation of species specific <b>diets</b> for exotics. The <a href="#">www.sugar-glider-store.com/insecteater-diet-canned-food-eng.html</a> · Standard (Grade: 4)	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)

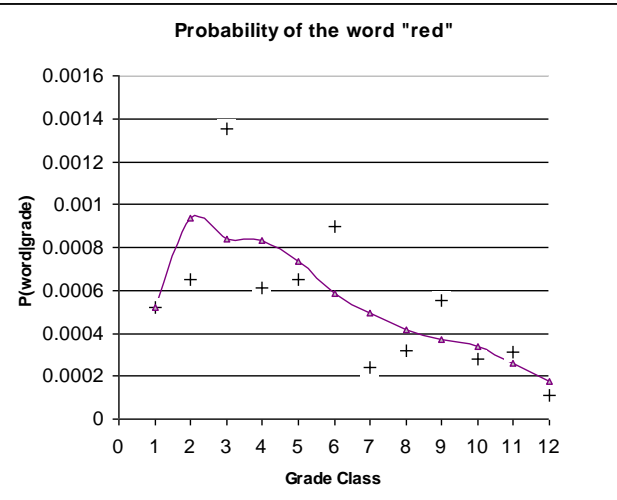
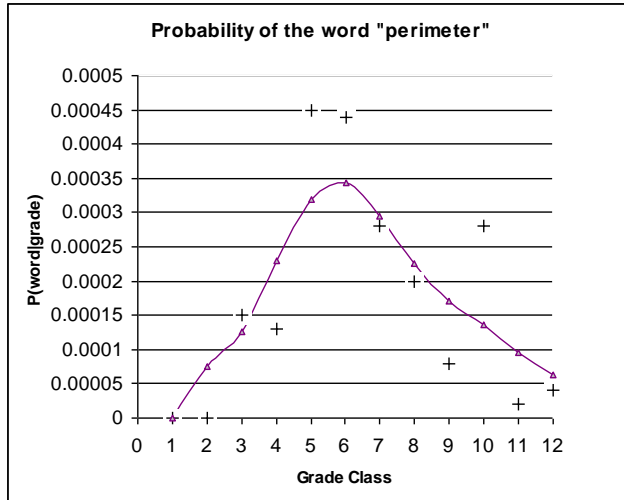
$$RG_{FK} = 0.39 \cdot [Words / Sentence] + 11.8 \cdot [Syllables / Word] - 15.59$$

- 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

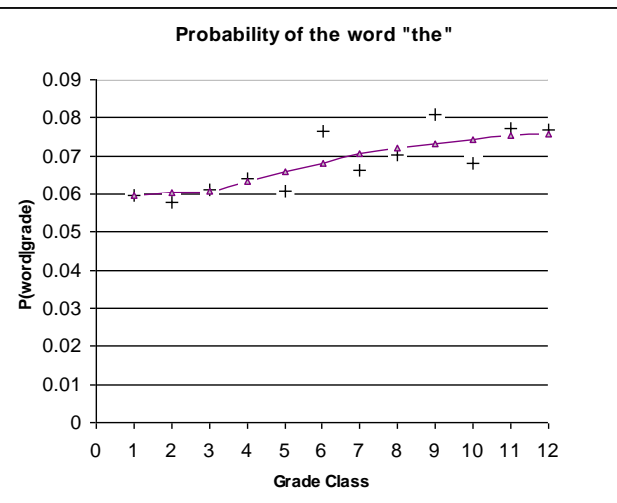
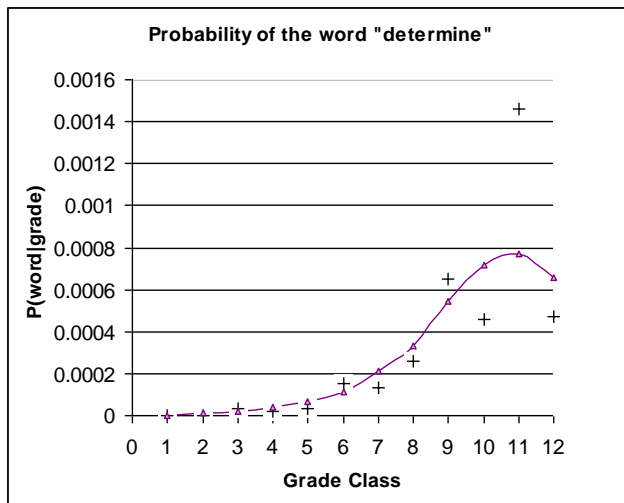
[Collins-Thompson & Callan: HLT 2004]

perimeter



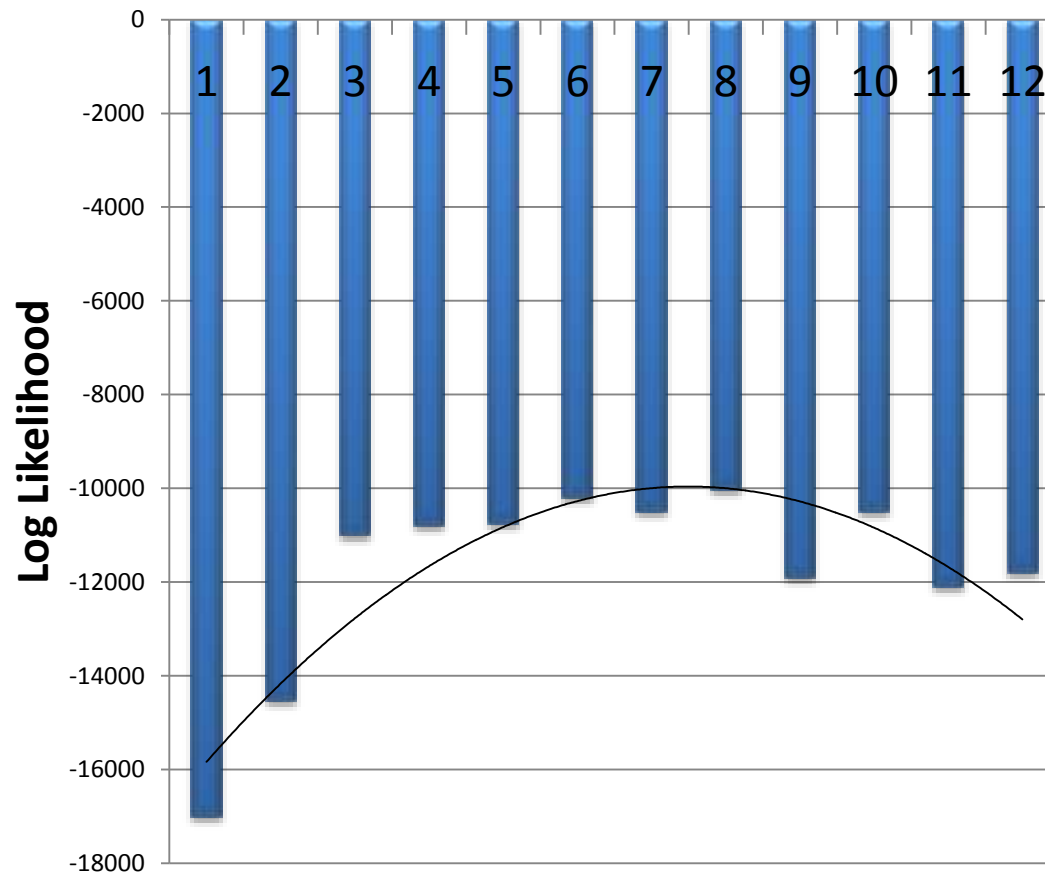
red

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 ( $\mu_w, \sigma_w$ ) 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

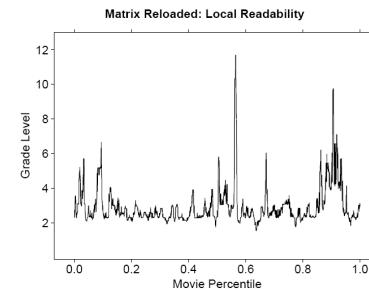
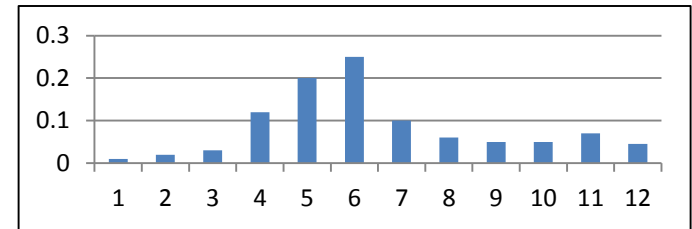
acidic	1.425
soda	1.425
acid	1.408
typical	1.379
angle	1.362
press	1.318
radio	1.284
flash	1.231
levels	1.229
pain	1.220

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)
- 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

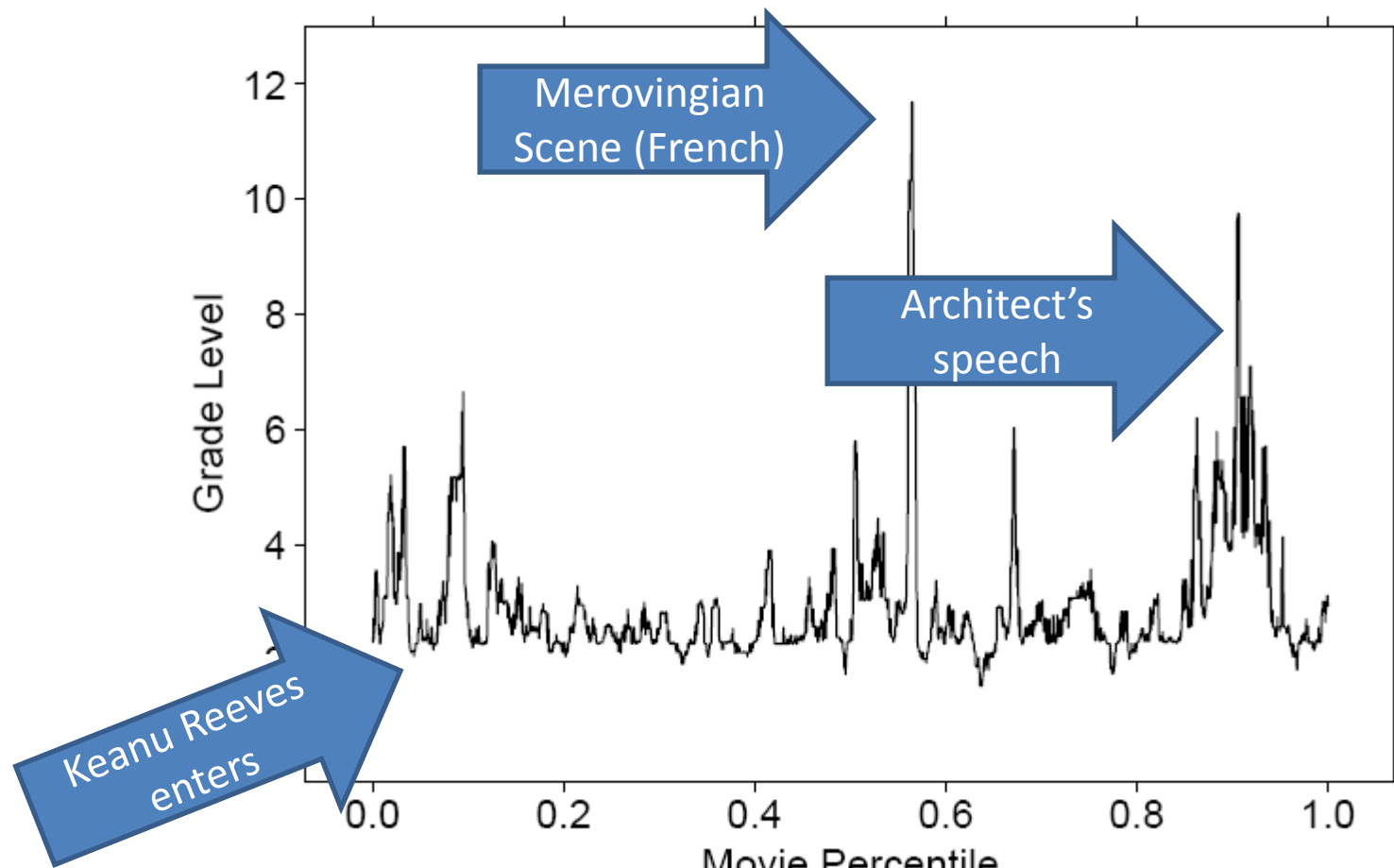


Health article: *Bronchitis, efficacy ...*

# Local readability within a document

Movie dialogue in “The Matrix: Reloaded”

**Matrix Reloaded: Local Readability**



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

Enriching the Web with Readability Metadata

# 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.

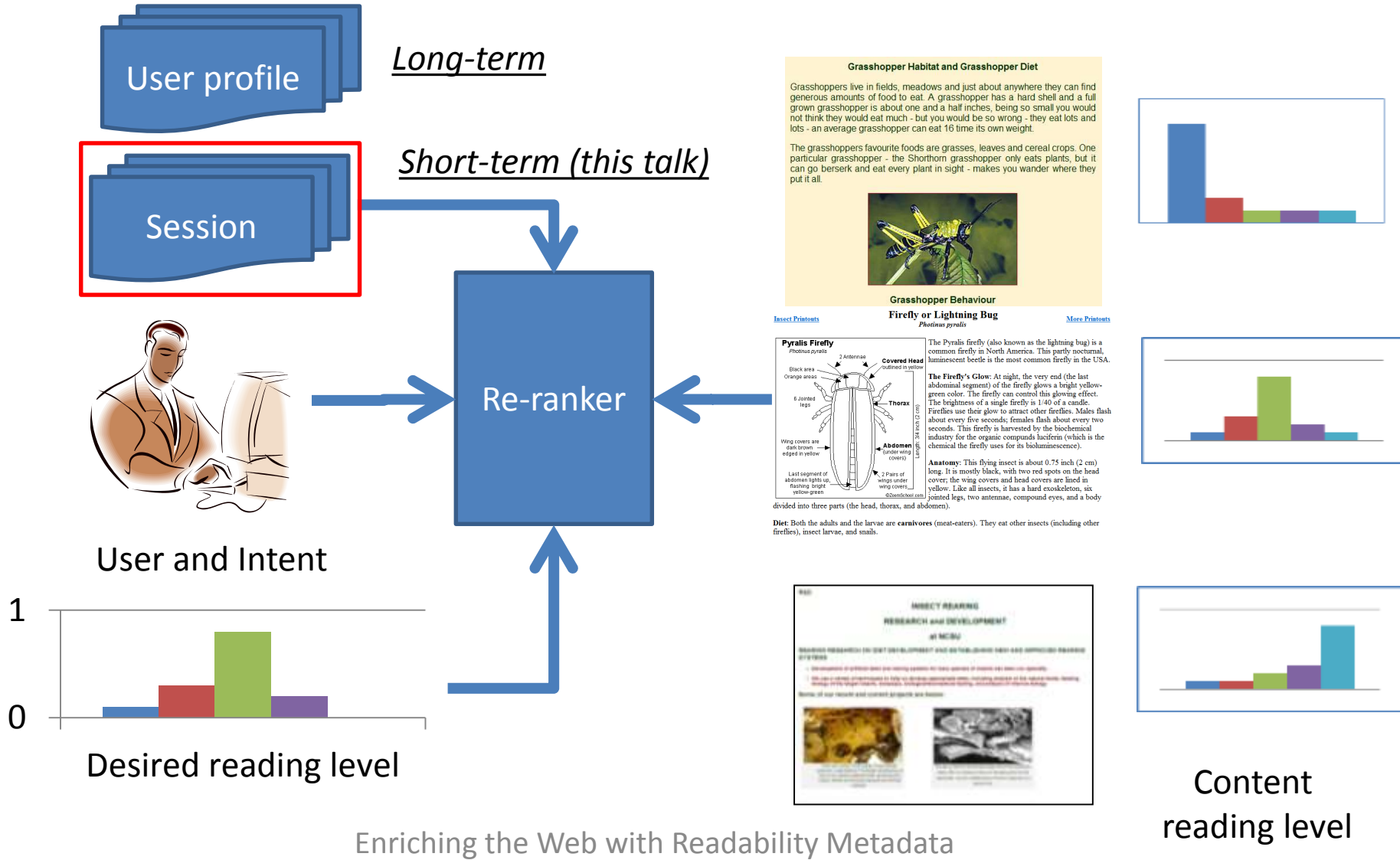
*A search result should be at the reading level  
the user wants for that query.*

Intent Models



Content Models

# Personalization by modeling users and content

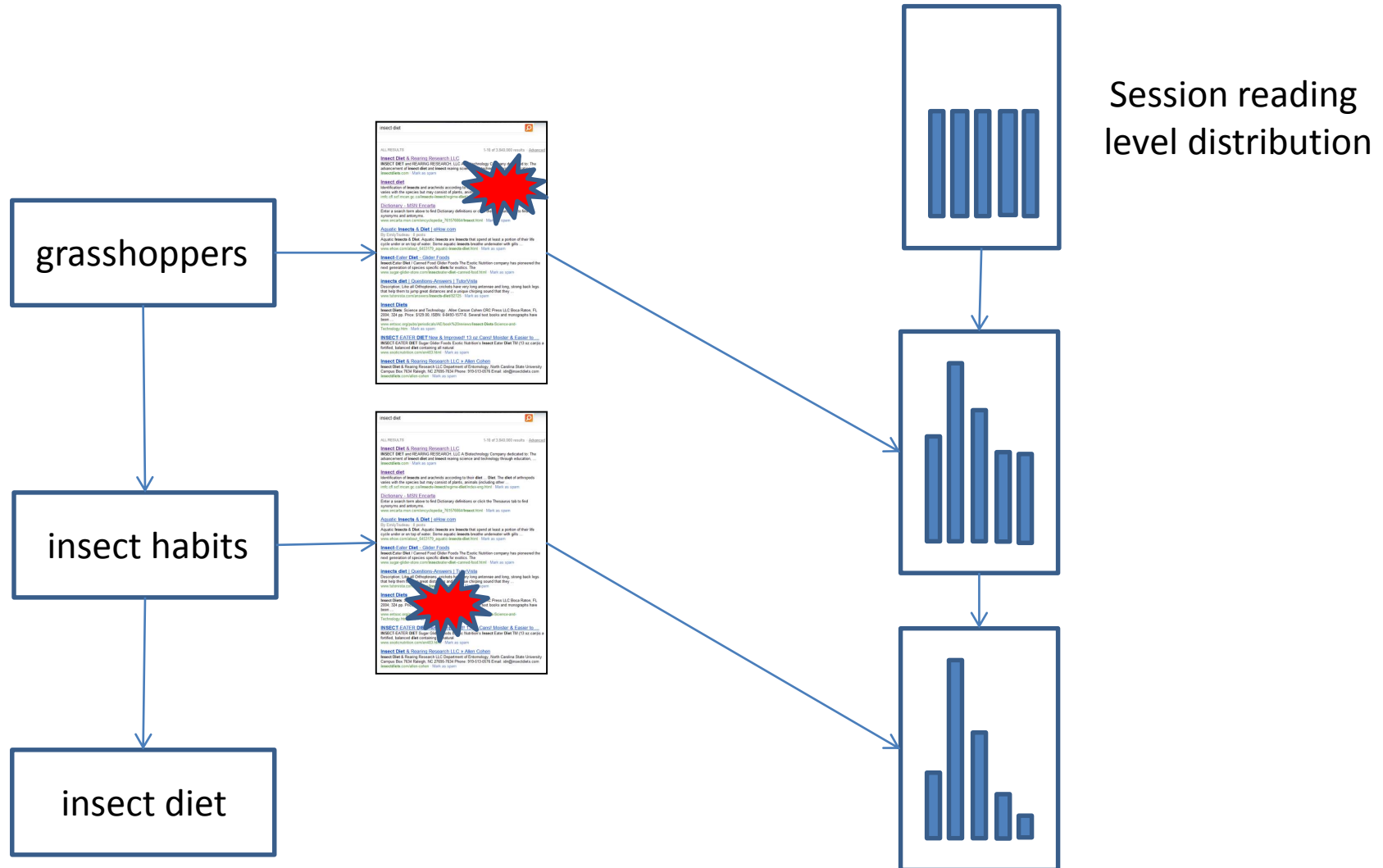


# How could a Web search engine personalize results by reading level?

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



# 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*

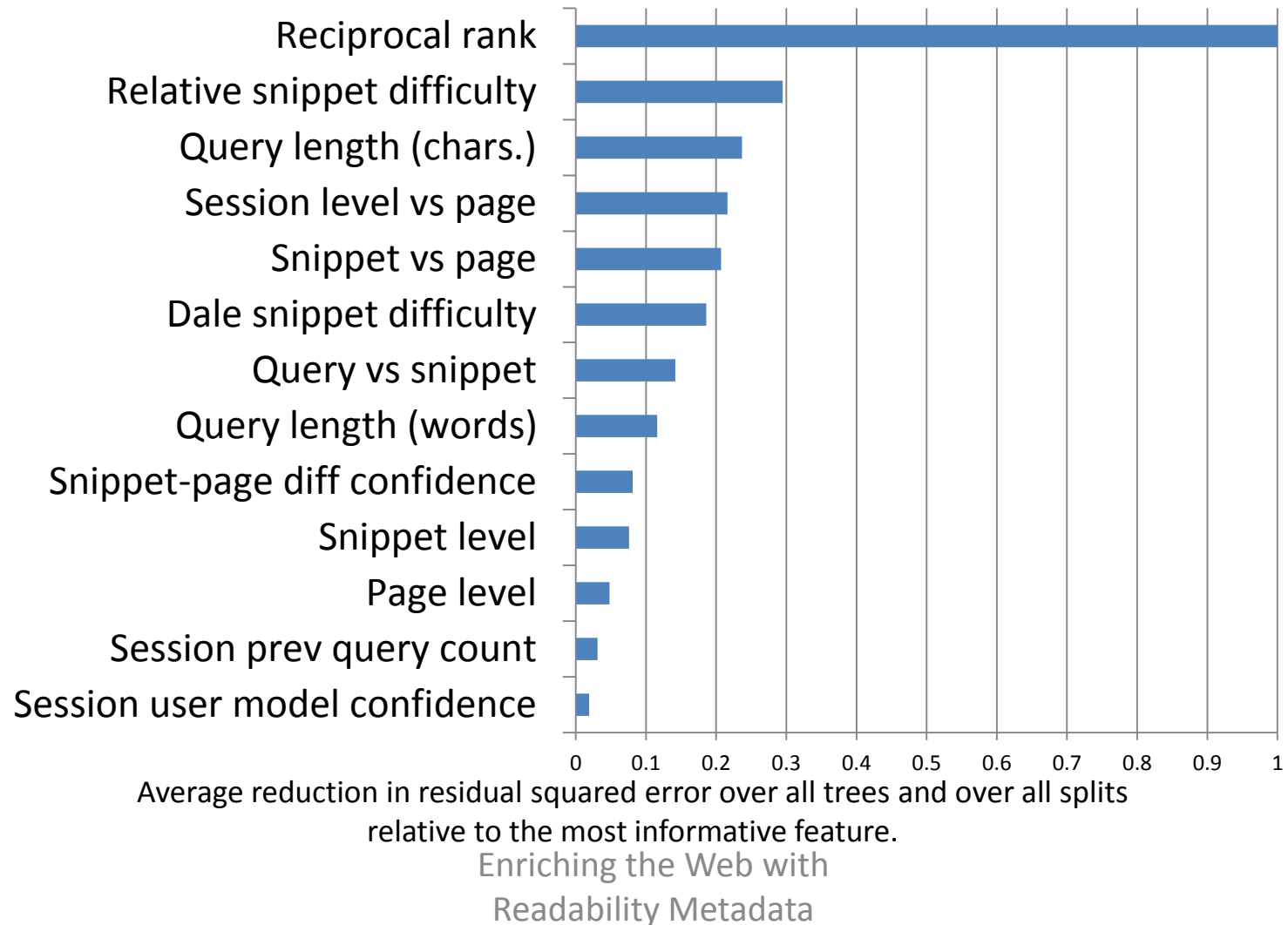
Point-Gain in Mean Reciprocal Rank of Last-SAT click

- 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:  $\text{Gain} \geq 1.0$  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?



# Application: Improving snippet quality

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

Snippet Difficulty: Medium

Click!

Welcome to IREaR

## INSECT REARING EDUCATION and RESEARCH PROGRAM Department of Entomology, North Carolina State University

Page Difficulty: High

Retreat!!

The IREaR program is dedicated to:

- The advancement of insect rearing science and technology
- The education of students and rearing professionals in the most up-to-date rearing practices
- The development of quality control and process control in rearing systems
- The recognition that insect rearing is a science and technology

### Insect Rearing Philosophy

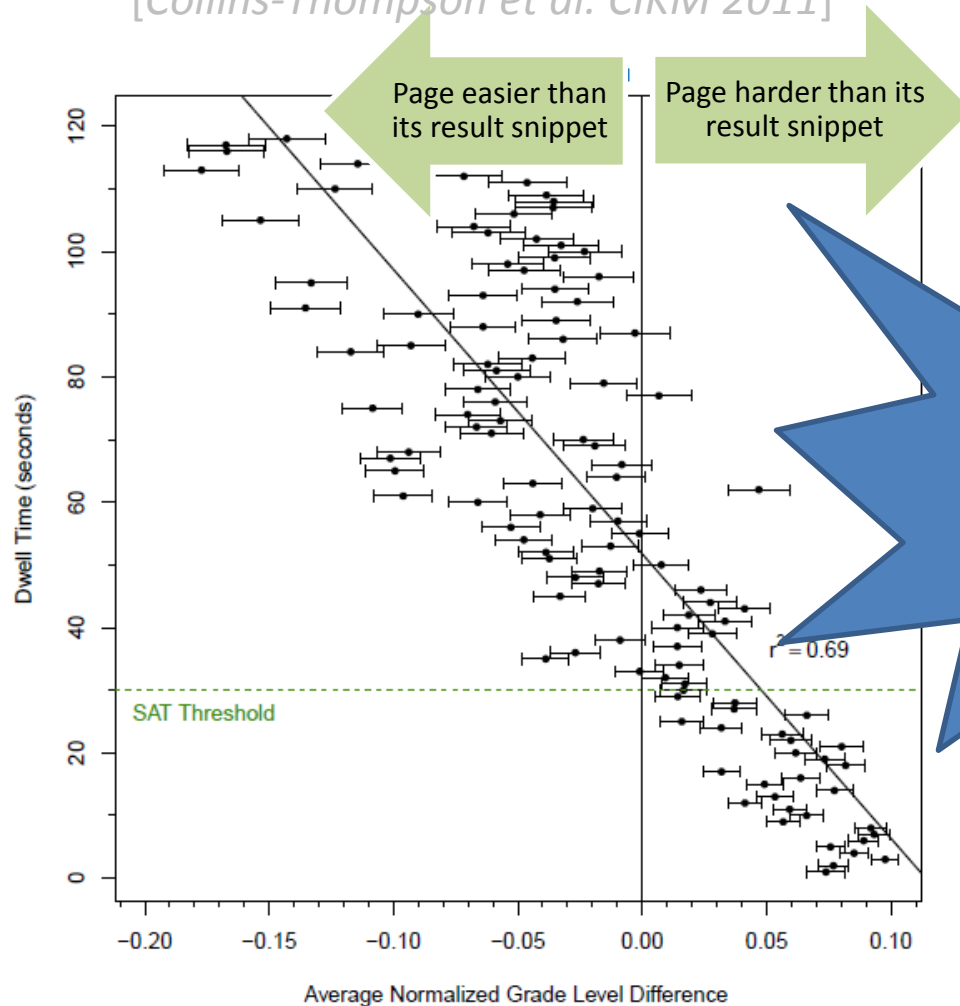
1. Rearing is treated as a science and technology
2. Knowing the insect thoroughly is essential to successful programs
3. Rearing is interdisciplinary
4. Good rearing systems must provide all the needs of insects (everything that insects need in nature must be provided in artificial rearing systems)

### Who Makes up the Insect Rearing Program at NCSU?

- The IREaR program is based in the NCSU Entomology Department, with extensive interactions and cooperation with other departments

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](http://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](http://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](http://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](http://www.thebigview.com/spacetime/quantumtheory.html)

## Top 4 results

# [quantum theory] + lower difficulty

## [Quantum Theory - PS3 - IGN - Sony PlayStation 3 ...](#)

[PlayStation 3](#) · [29 photos](#) · [Walkthroughs](#) · [Cheats](#)

Sep 28, 2010 · **Quantum Theory** is a game whose design is dated despite being a week old. It's a game that feels like it didn't ...

[ps3.ign.com/objects/142/14288075.html](http://ps3.ign.com/objects/142/14288075.html)

**IGN**

2.5/10

score

## [Quantum Theory : Mix That Drink](#)

I wonder where the **Quantum Theory** cocktail got its name. There's nothing incomprehensible about this cocktail, and it's not as mind-blowing as, say, the Zombie.

[mixthatdrink.com/quantum-theory](http://mixthatdrink.com/quantum-theory)

## [Quantum Theory Cheats, Codes, and Secrets for PlayStation 3 - GameFAQs](#)

For **Quantum Theory** on the PlayStation 3, GameFAQs has 51 cheat codes and secrets.

[www.gamefaqs.com/ps3/954470-quantum-theory/cheats](http://www.gamefaqs.com/ps3/954470-quantum-theory/cheats)

## [Quantum Theory Cheats - Playstation 3 - ActionTrip -- What we lack ...](#)

This page offers the most up-to-date **Quantum Theory** Playstation 3 cheats, codes, and hints. Besides our impressive collection of **Quantum Theory** and other cheats, ...

[www.actiontrip.com/cheats/ps3/quantum-theory.phtml](http://www.actiontrip.com/cheats/ps3/quantum-theory.phtml)

## 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](http://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...](http://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](http://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](http://www.cinderella.dk/index.htm)

## [Cinderella - School of Ballroom Dancing](#)

About Us: Home | Contact Us : Welcome 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](http://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](http://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](http://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](https://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](http://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:

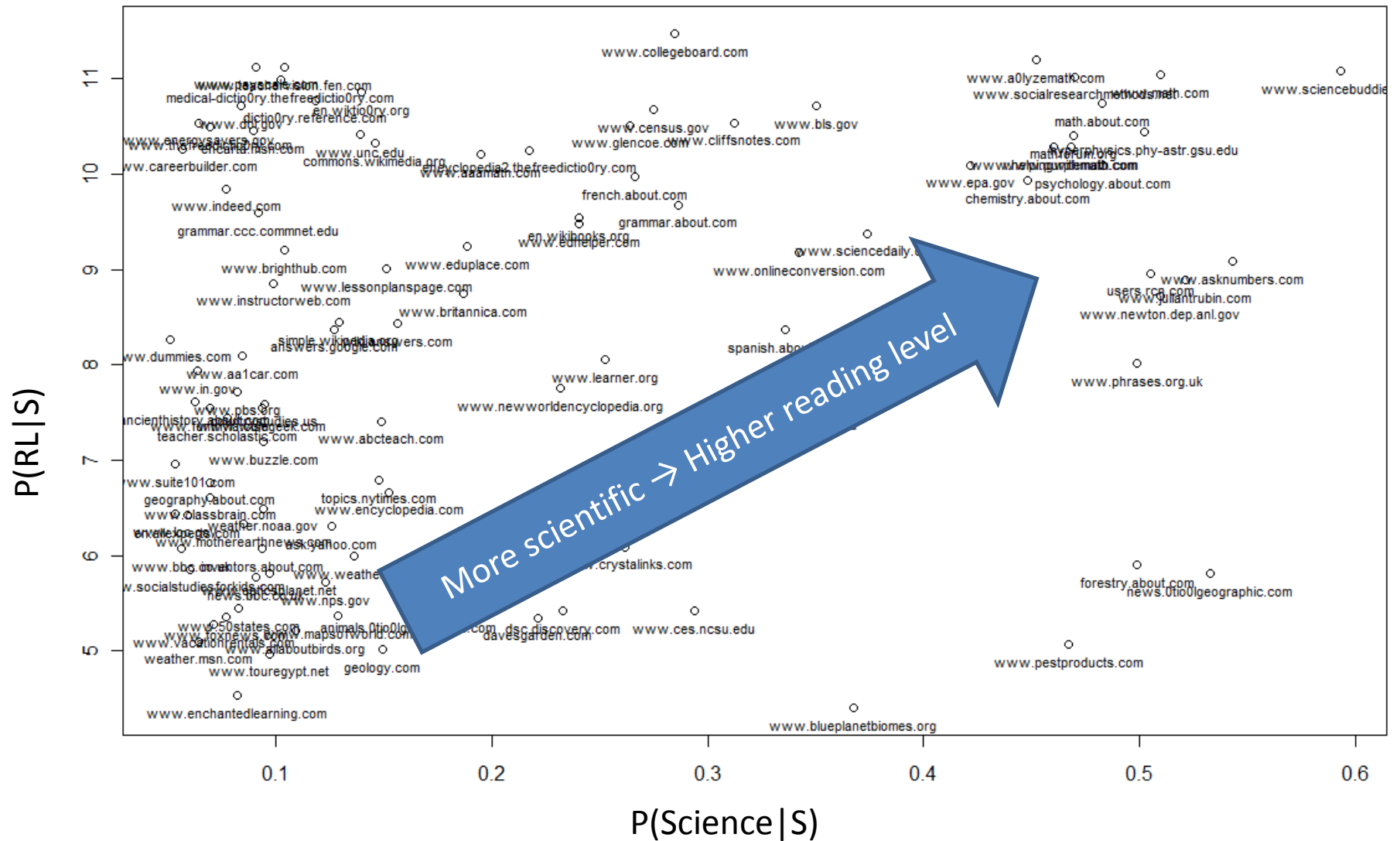
We have initially operated BAMBI as a total power receiver for several

# 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

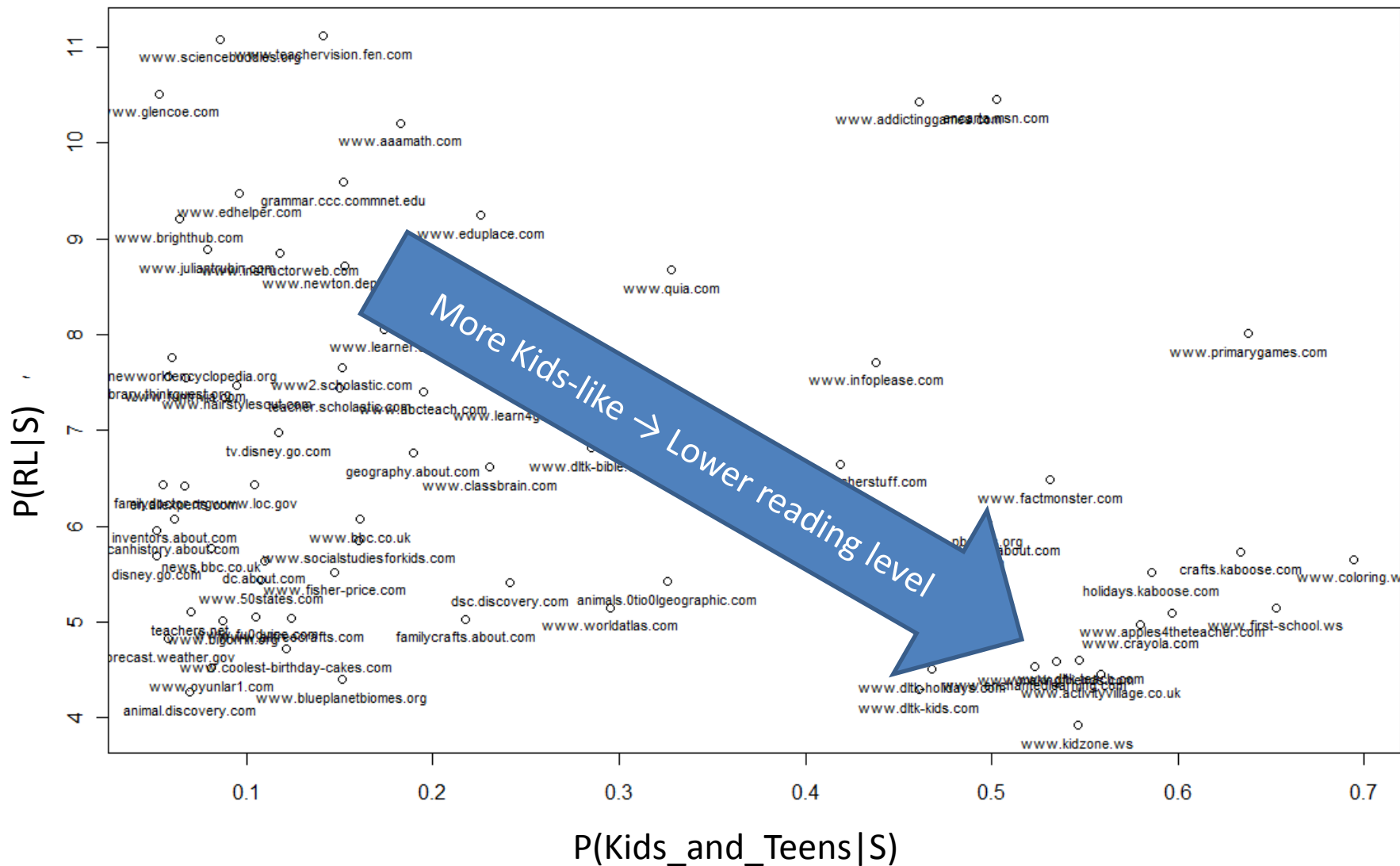


# $P(RL|S)$ against $P(\text{Science}|S)$



Enriching the Web with Readability Metadata

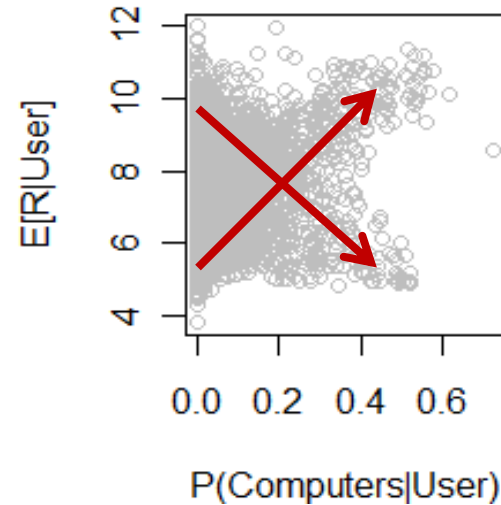
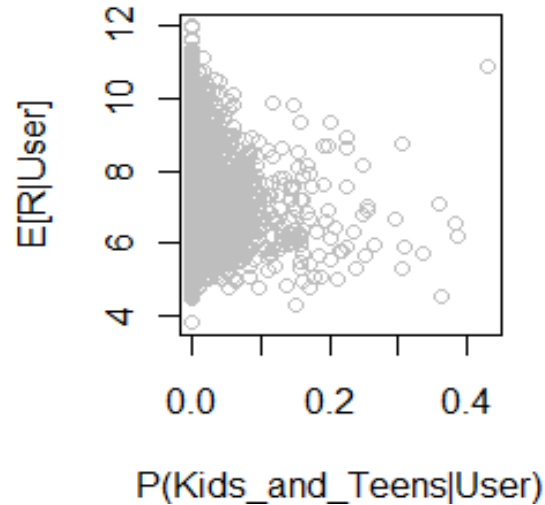
# $P(\text{RL}|\text{S})$ against $P(\text{Kids\_and\_Teens}|\text{S})$



Enriching the Web with Readability Metadata

# 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 together 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,  $H(RL|S) > 2$

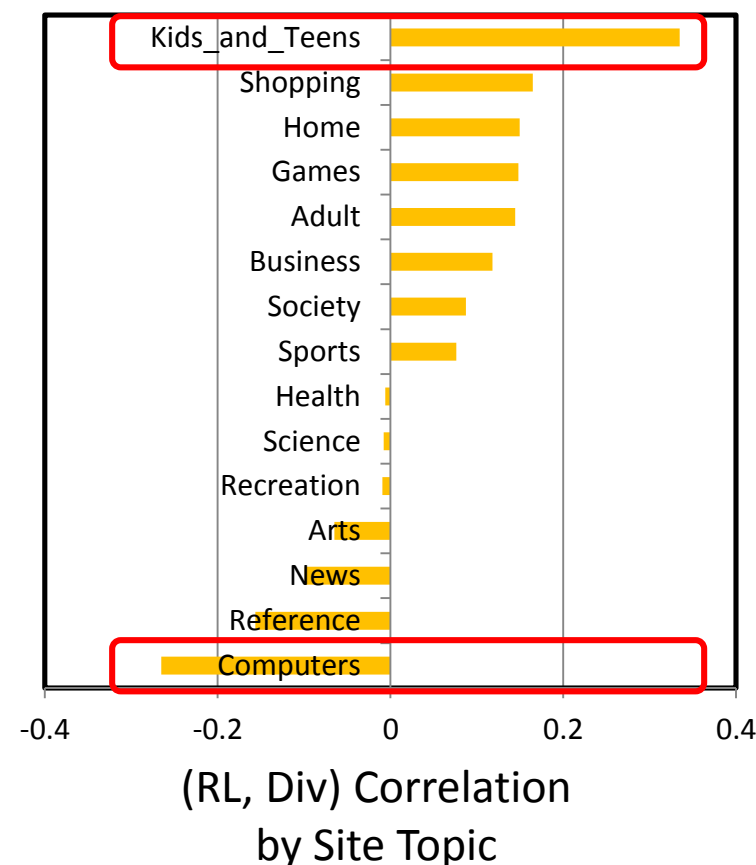
Website	H(RL S)	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	Count	E(RL S)
www.dltk-kids.com	2.02	0	0	0.2	0.5	0.2	0.1	0	0	0	0	0	0	39	4.4
www.dltk-teach.com	2.1	0	0	0.2	0.4	0.2	0.2	0	0	0	0	0	0	26	4.47
www.dltk-holidays.com	2.07	0	0	0.2	0.5	0.1	0	0.1	0	0	0	0	0	31	4.65
psychology.about.com	2.32	0	0	0	0	0	0	0.1	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		
	$\text{Div}_R(U s)$	$\text{Div}_T(U s)$	$\text{Div}_{RT}(U s)$
$E[R s]$	0.052	0.081	0.095

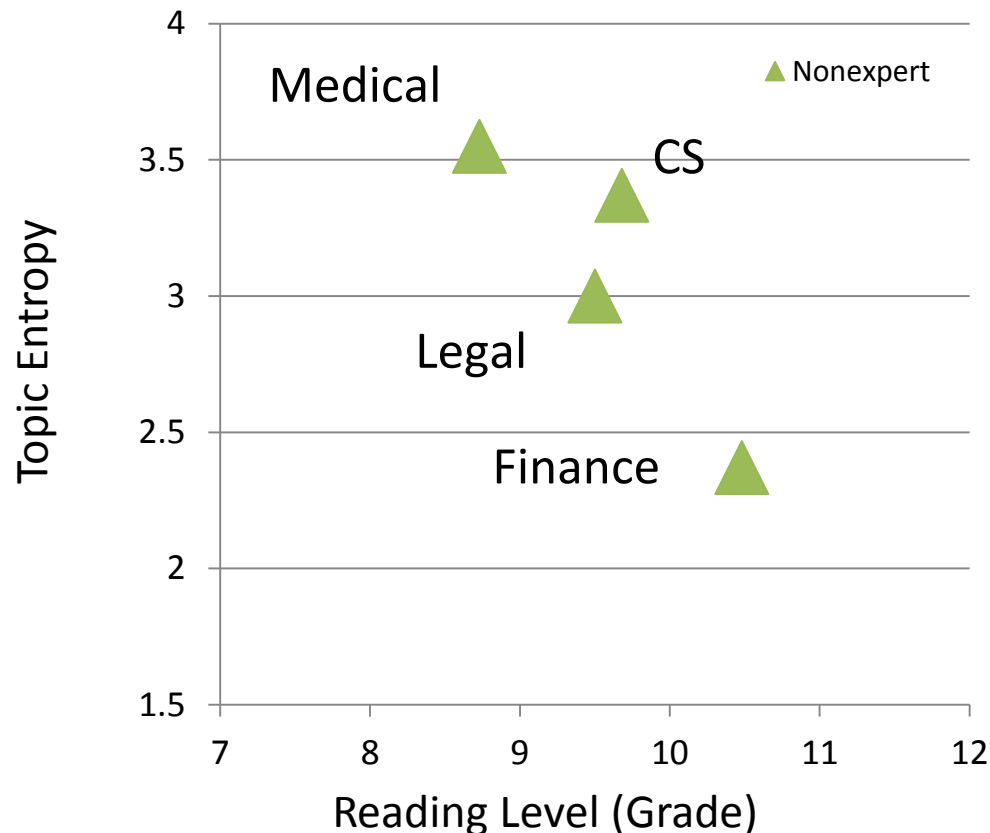
- ▶ ...But a breakdown of sites by topic reveals stronger relationships
  - ▶ Computer sites with high reading level attract focused visitors
  - ▶ Kids sites with high reading level attract diverse visitors





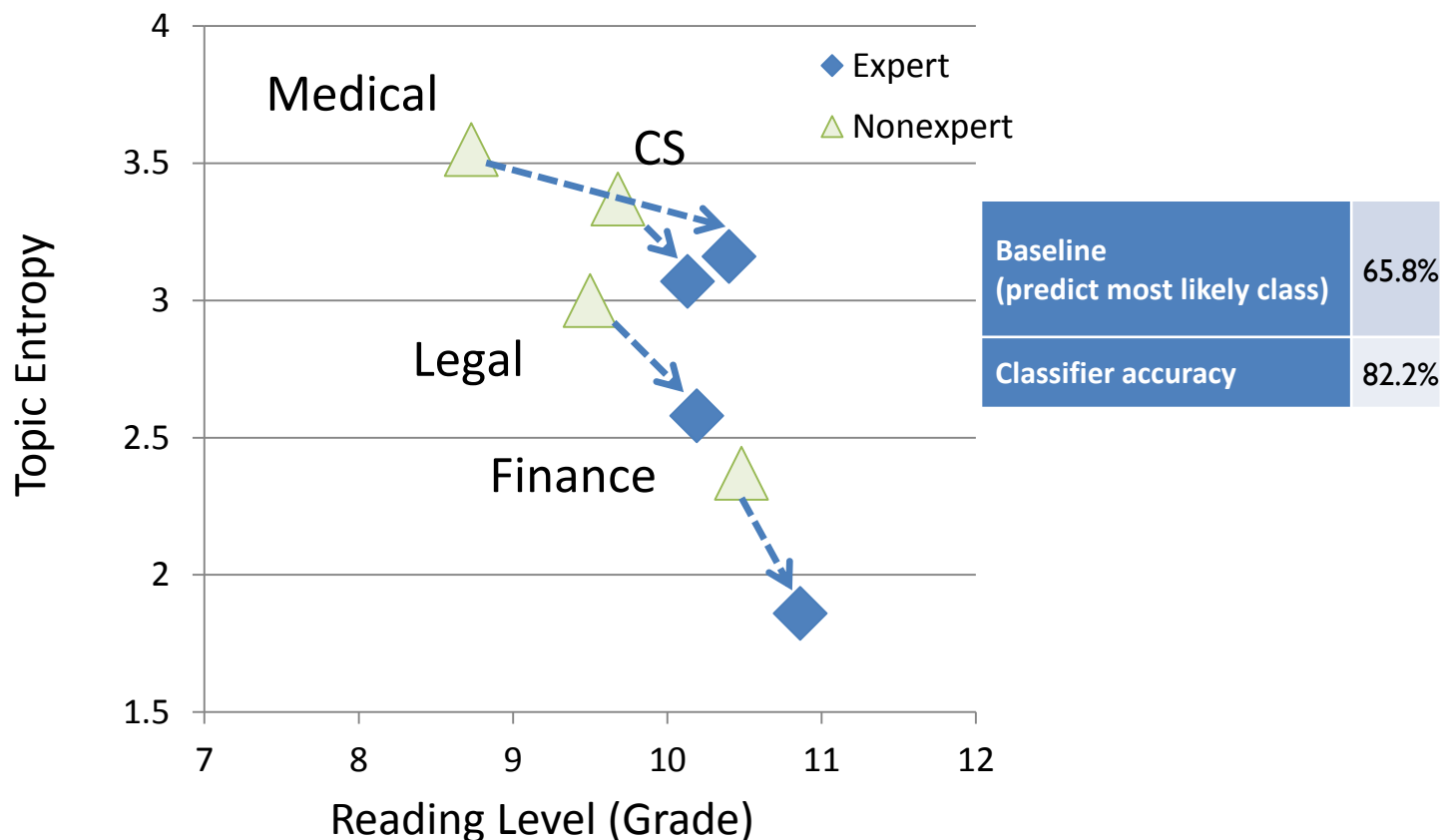
# Reading level and topic entropy features can help separate expert from non-expert websites

*[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]*



# 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
<a href="http://socialsecurity.gov">socialsecurity.gov</a>	Government retirement/disability
<a href="http://collegeboard.com">collegeboard.com</a>	Entrance exam preparation, college application help
<a href="http://softwarepatch.com">softwarepatch.com</a>	Find software patches
<a href="http://fileinfo.com">fileinfo.com</a>	Find programs to open file types
<a href="http://msdn.microsoft.com">msdn.microsoft.com</a>	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
$$E[R | \text{Session}] - E[R | \text{User}] > 4 \text{ grade levels}$$
- 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
Medical tests	tests	2.22
	test	1.99
	sample	1.94
College entrance	digital	1.88
	options	1.87
	aid	1.87
Financial aid	effects	1.84
	education	1.77
	forms	1.76
Gov’t forms	plan	1.74
Job search	pay	1.71
	medical	1.69
	learning	1.62

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

Enriching the Web with Readability Metadata

‘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			
Medical tests College entrance		tests	2.22	best	-0.42	Shopping! Exploration Leisure		
		test	1.99	football	-0.45			
		sample	1.94	store	-0.46			
		digital	1.88	great	-0.47			
		options	1.87	items	-0.52			
Financial aid		aid	1.87	new	-0.53			
		effects	1.84	sale	-0.61			
		education	1.77	games	-0.65			
	Gov't forms		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			

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

Enriching the Web with Readability Metadata

‘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	
Top word	Log ratio	Title word	Log ratio
Medical tests	2.7	Shopping!	-0.81
College entrance test sample		Exploration	
Financial aid application		Leisure	
Gov't forms			
plan	1.7	food	-0.81
pay	1.7	news	-0.82
medical	1.69	music	-1.02
learning	1.62	all	-1.35

Future work:

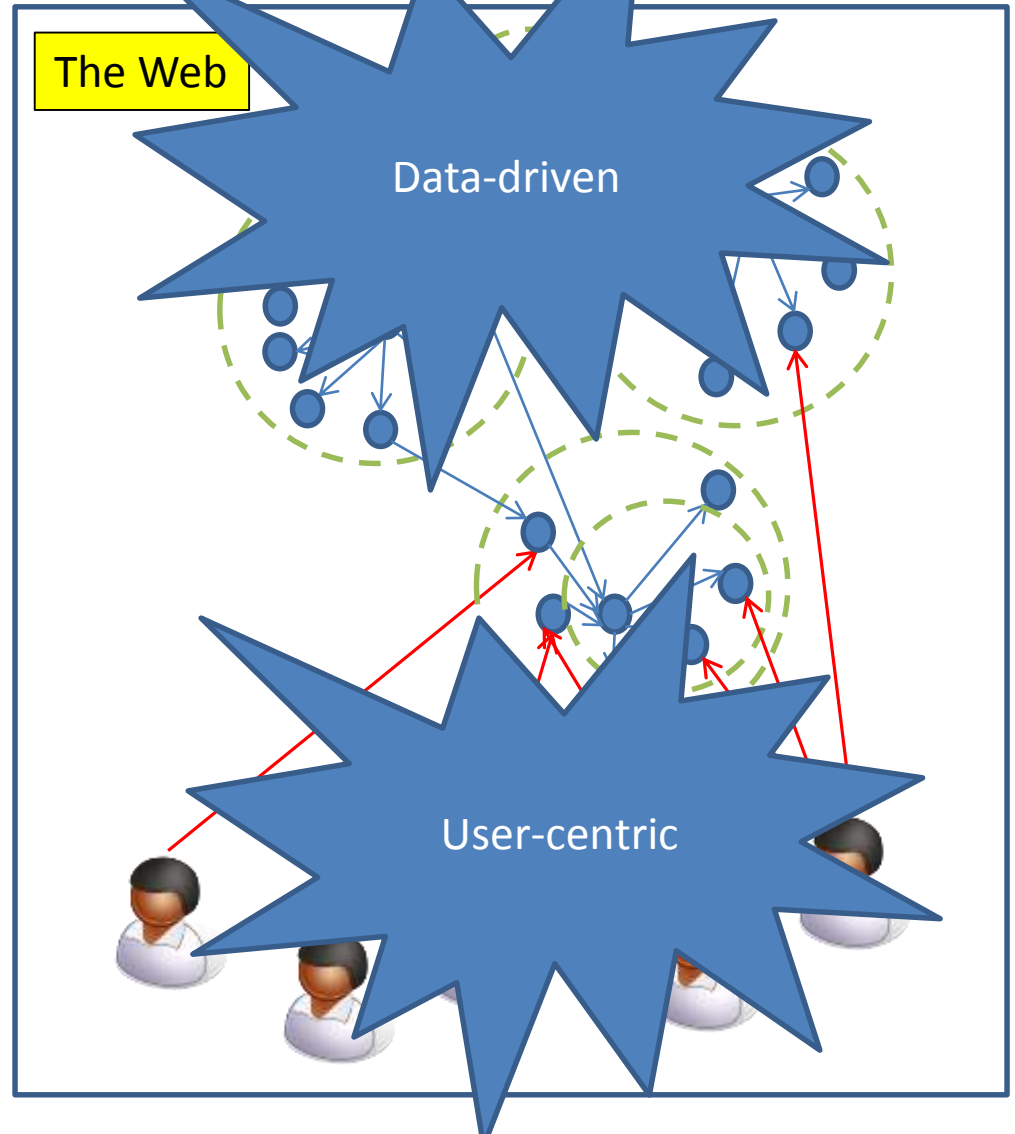
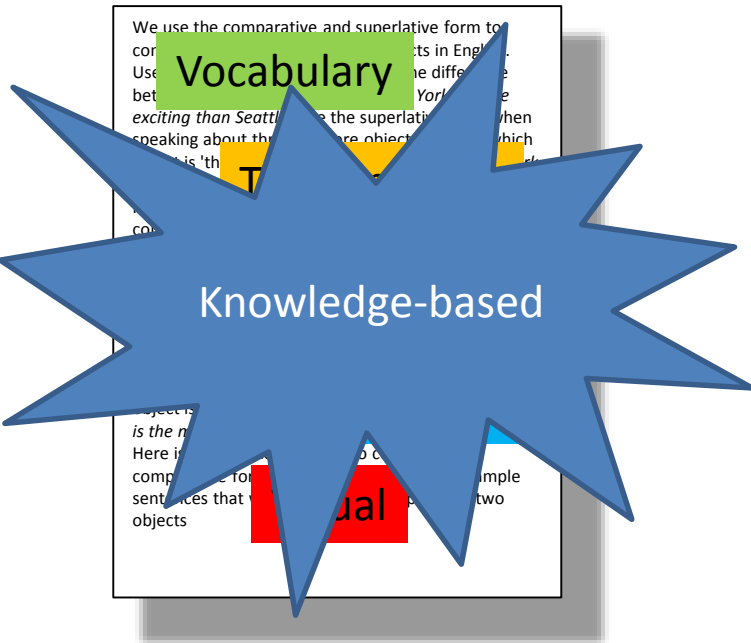
1. Identify & predict stretch tasks
2. Decide how and when to provide support
3. Determine helpful background or alternatives

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

Enriching the Web with Readability Metadata



# 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

- Analyzing movie scripts with Keanu Reeves dialogue



- Data-driven, personalized readability measures
- Adapting content to users
  - Enrich, augment, rewrite
- Adapting users to content
- Influencing search presentation and interaction

- Web-scale speed and reliability
- Exploiting new content forms
  - Blogs, wiki structure & edits
- Adapting to different tasks and populations
- Human computation/crowdsourcing
- Predicting quality/authority



Relevance for applications

Enriching the Web with Readability Metadata

# 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](mailto:kevynct@umich.edu)

Web site:

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