Enhancing document representations using analysis of content difficulty

Models, Applications, and Insights

Kevyn Collins-Thompson

Associate Professor
University of Michigan

School of Information &
College of Engineering, EECS

Thanks to my valued collaborators:

<u>Rohail Syed</u>	*Univ. of Michigan*
<u>Karthik Raman</u>	*Cornell / Google Research*
Gwen Frishkoff	*University of Oregon*
<u>Elliot Schumacher</u>	*Carnegie Mellon / JHU*
Jamie Callan	*Carnegie Mellon*
Maxine Eskenazi	*Carnegie Mellon*
Sebastian de la Chica	*Microsoft*
Paul Bennett	*Microsoft Research*
Susan Dumais	*Microsoft Research*
Ryen White	*Microsoft Research*
Michael Heilman	*Educational Testing Service*
<u>Paul Kidwell</u>	*Purdue / LLNL*
Guy Lebanon	*Georgia Tech / Google*
David Sontag	*MIT*

And the KDD Document Intelligence Workshop organizers

Documents on a topic can 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

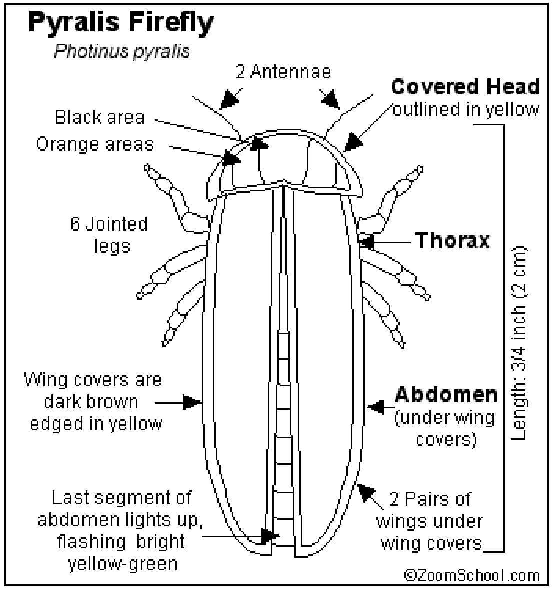
Query [insect diet]: Lower difficulty

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

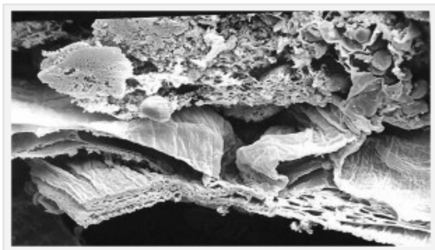
at NCSU

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

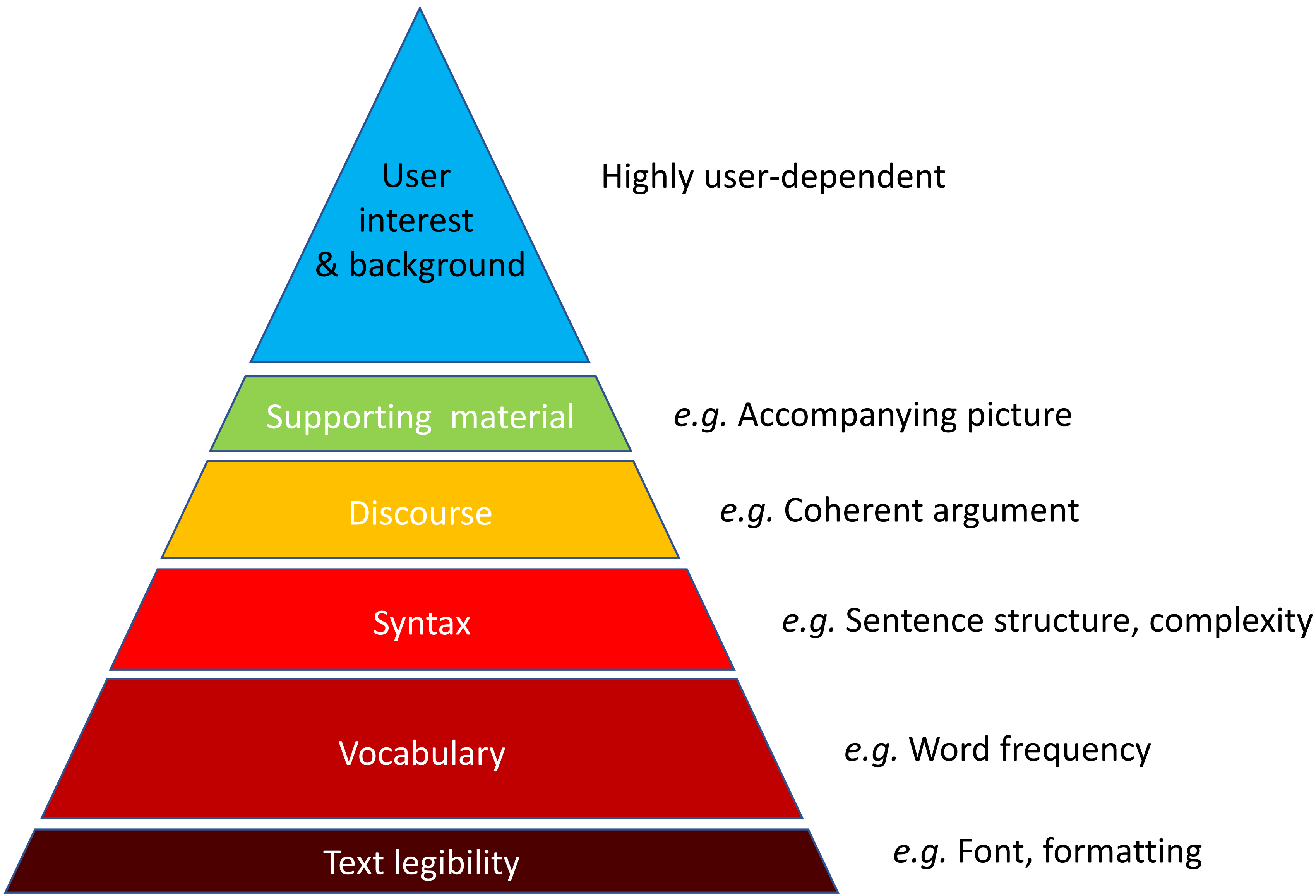
- Students at different grade levels
- Non-native speakers
- General population
  - Overall language proficiency or literacy
  - Familiarity or expertise in specific topic areas

Long-term goal: Optimally connecting users & information for personalized learning and discovery

1. How can we create richer representations of user and content for supporting learning (at Web scale)?
2. How can we integrate models of human learning and cognition into search and recommender algorithm objectives, features, and evaluation?
3. How can search engine algorithms support important educational goals like robust long-term retention, or increase in curiosity…not just short-term learning?

This talk: Analyzing content difficulty enhances document representations for understanding and supporting readers.

What makes text difficult to read and understand?



Traditional readability measures don't work for Web and other non-traditional 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 and other content is often short, noisy, less structured
  - Page body, titles, snippets, queries, captions, …
  - Health questionnaires, surveys
- Billions of pages → computational constraints on approaches
- We focus on generative *vocabulary-based prediction models* that learn fine-grained models of word usage from labeled texts

Model 1: Vocabulary-based generative models: smoothed unigram language models

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

1. Model each grade G_i as a word histogram θ_i.
2. Smooth θ_i by combining evidence from nearby grades.

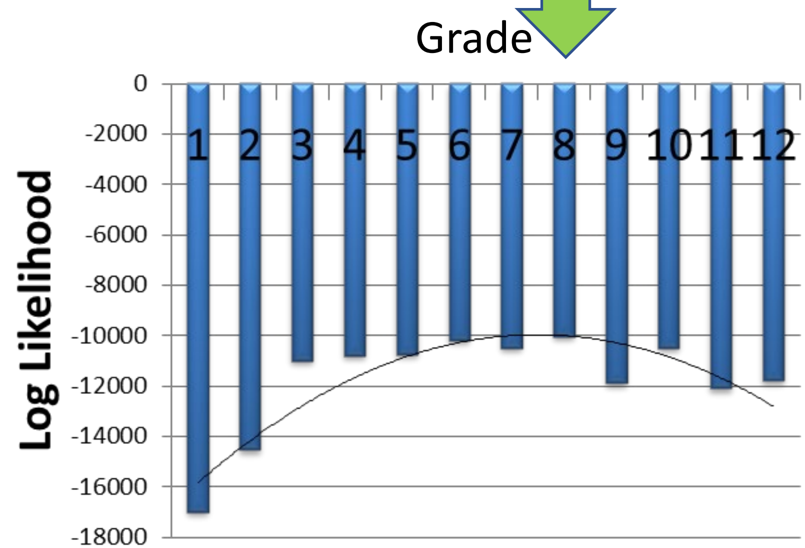
$$\hat{\theta}_i(w) = \sum_{j=1\ldots|G|} \phi_h(i,j)\hat{P}(w|G_j)$$

3. Compute the likelihood of the text in each grade model θ_i .

$$\log P(G_i|T) = \sum_{w \in T} C(w) \log \hat{\theta}_i(w) + \log Q$$

4. <u>Prediction</u>: Select the **most likely** grade.

Type	*Grade 1* $P(w \mid G_1)$	*Grade 5* $P(w \mid G_5)$	*Grade 12* $P(w \mid G_{12})$
the	0.080	0.090	0.100
a	0.060	0.050	0.060
red	0.020	0.005	0.0007
ball	0.010	0.0001	0.00005
was	0.010	0.010	0.200
perimeter	0.003	0.04[illegible]	0.004
optimized	0.00001	0.00[illegible]	0.010



Grade 8 document: 1500 words

Which words are most 'distinctive' of each grade in these language models?

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

*These values are computed using a Fisher information-type statistic

Model 2: Estimate word acquisition events

[Kidwell, Lebanon, Collins-Thompson EMNLP 2009, J. Am. Stats 2011]

- Inspiration: Dale-Chall list of 3000 words familiar to 80% of 4^{th} -graders.

- Documents can contain high-level words but still be low–level, e.g. teaching new concepts

- *(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

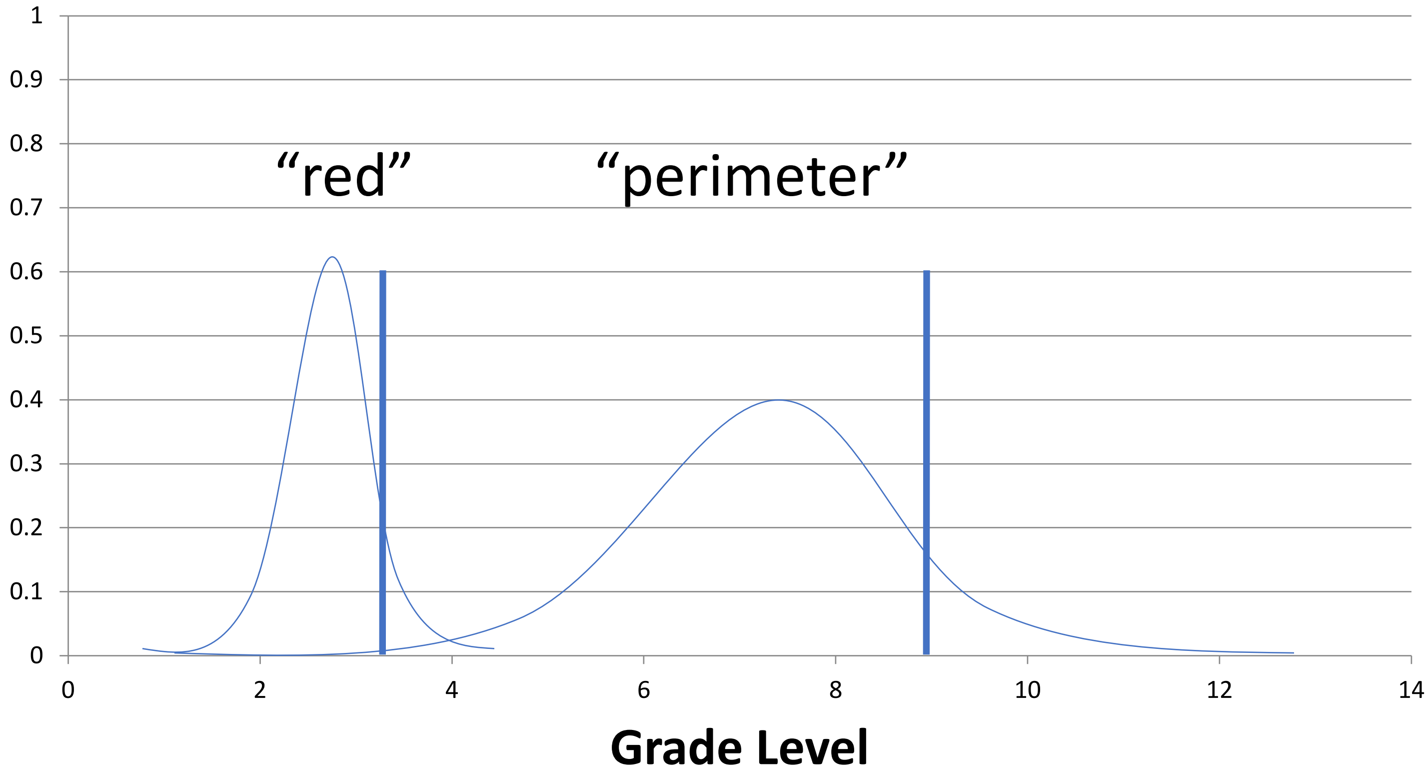
    A document d is readable at level t if s percent of the words in d are familiar at grade t.

- When does someone learn to read word w ?
  - Average acquisition age μ_w with standard deviation σ_w (Gaussian)
  - Fit Gaussian (μ_w, σ_w) parameters for each word *w*
  - Learn all parameters by maximum likelihood from labeled documents
- (*r*, *s*) parameters allow tuning the model for different scenarios

The *r* parameter controls the familiarity threshold for words

Level quantile for word *w*: $q_w(r)$

$q_{RED}(0.80) = 3.5$ $q_{PERIMETER}(0.80) = 8.2$

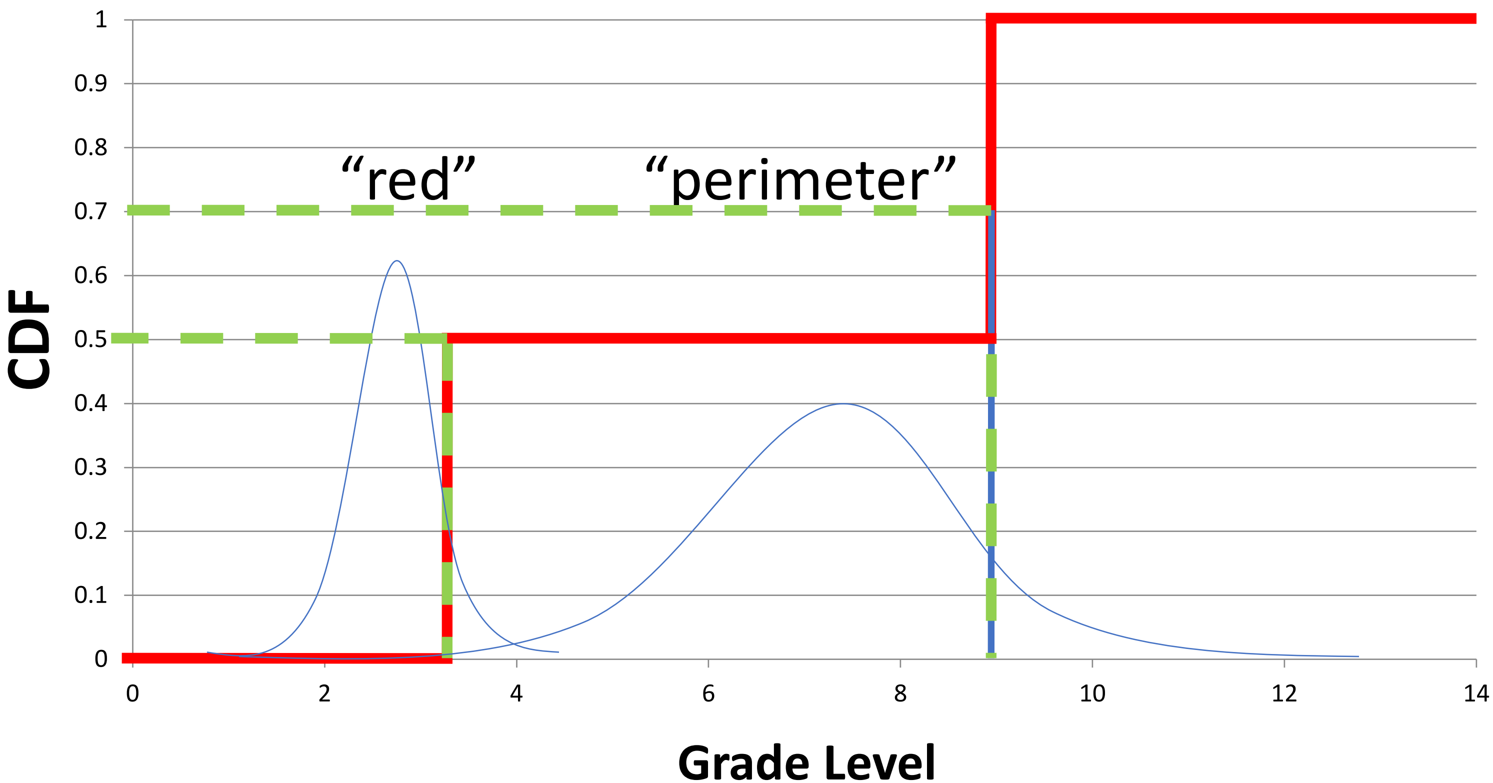


The *s* parameter controls required document coverage

Suppose: p(“red” | d) = p(“perimeter” | d) = 0.5

Predicted grade with s = 0.70: 8.8

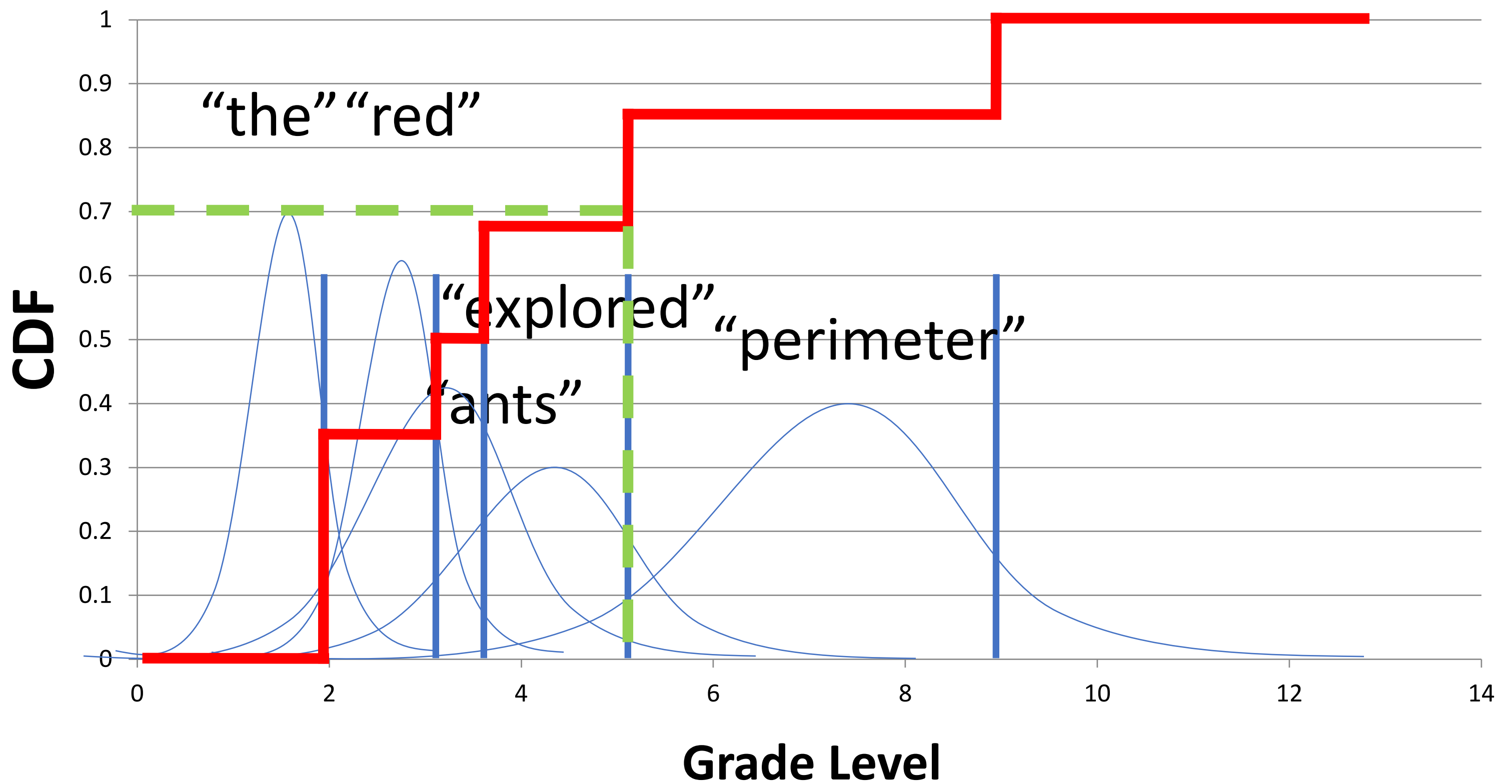
Predicted grade with s = 0.50: 3.5



Multiple-word example

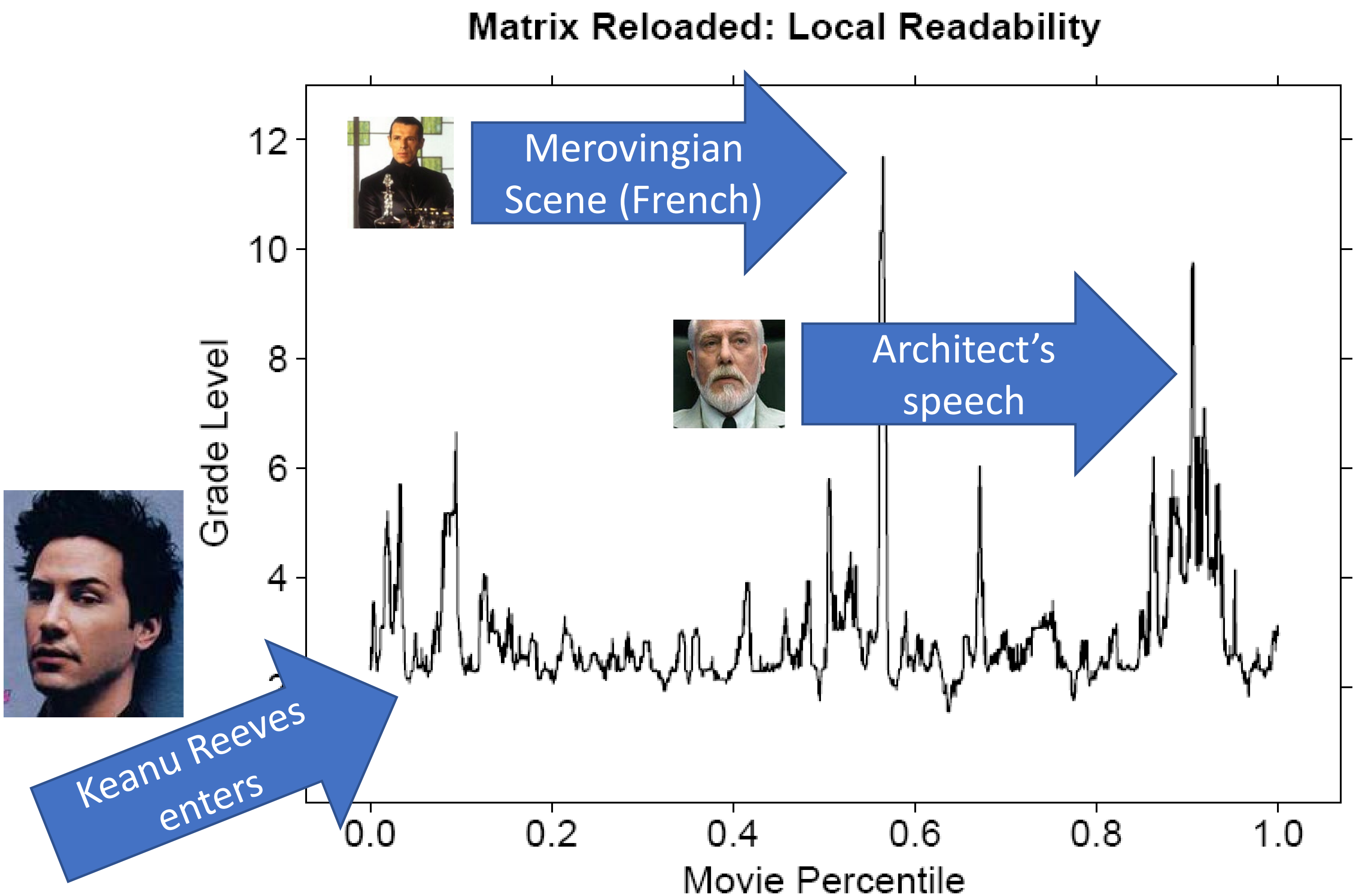
"The red ants explored the perimeter."

Predicted grade with s = 0.70: 5.3



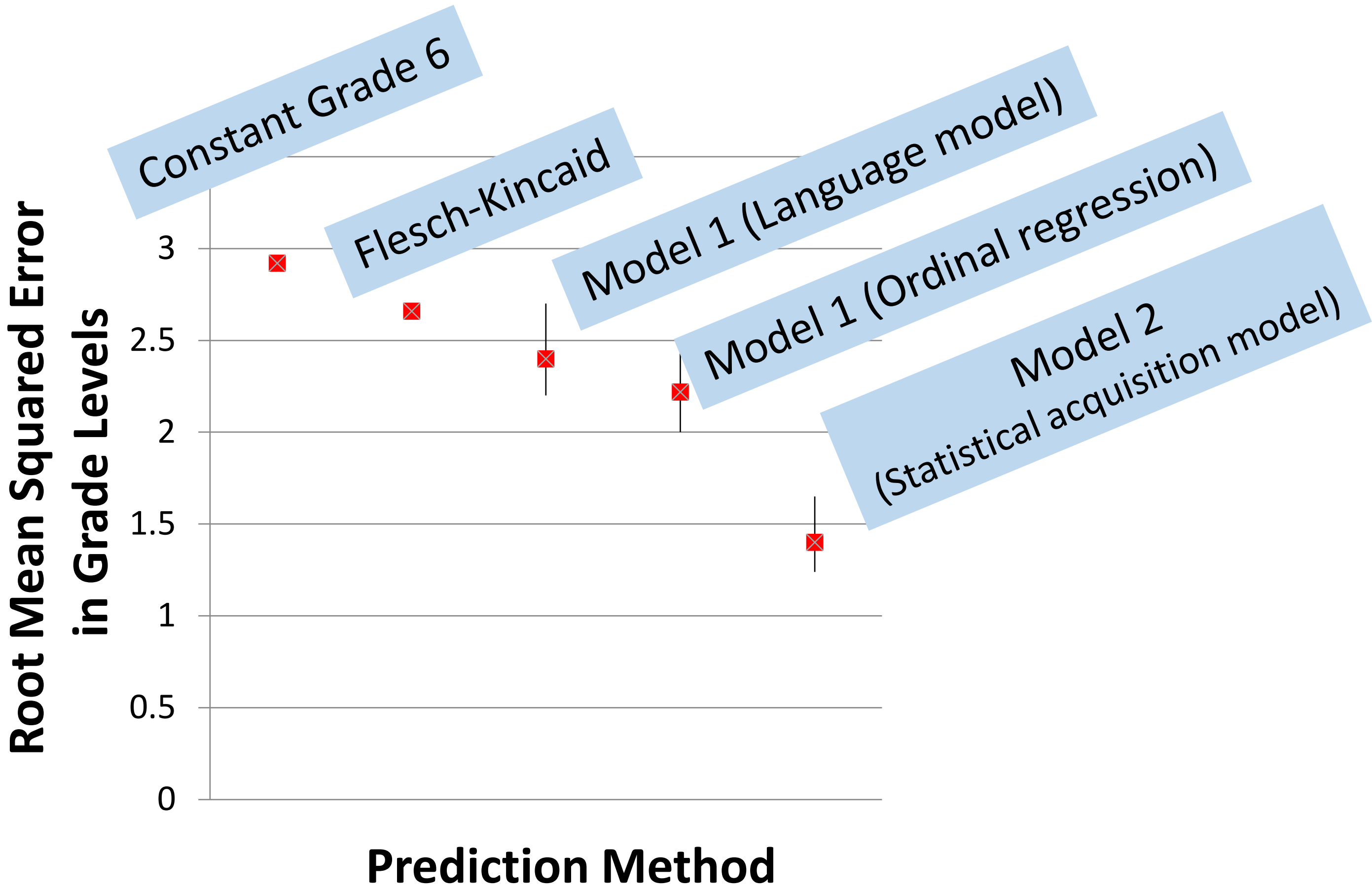
Local readability for documents with varying difficulty

Movie dialogue in "**The Matrix: Reloaded**"



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

More detailed vocabulary models improve prediction accuracy for Web content (lower is better)



Selected extensions for readability prediction methods I've explored

- First- *vs* second-language learners

  [M. Heilman, **K. Collins-Thompson**, J. Callan and M. Eskenazi. HLT 2007]

- Rich feature spaces: vocabulary, syntax

  [M. Heilman, **K. Collins-Thompson** and M. Eskenazi. ACL BEA workshop 2008]

- Crowdsourcing reliable difficulty labels

  [X. Chen, P.N. Bennett, **K. Collins-Thompson**, E. Horvitz. WSDM 2013]

- Single-sentence readability (relative difficulty)

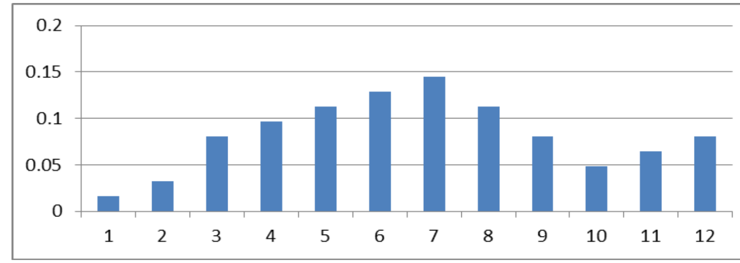
  [E. Schumacher, M. Eskenazi, G. Frishkoff, **K. Collins-Thompson**. EMNLP 2016]

See my computational readability survey linked on my UMichigan homepage
International Journal of Applied Linguistics 165:2

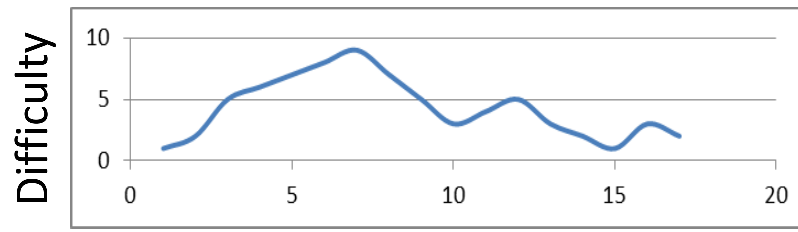
Reading difficulty provides a rich new representation of documents, sites, and users

<u>Level 1:</u> Documents

- Distribution over levels
- Can find key ‘words to learn’
- View change over time/ position



Health article: *Bronchitis, efficacy …*



<u>Level 2:</u> Web sites

- e.g. Reading level mean and variance across site pages

<u>Level 3:</u> User profiles

- Reading level via user clicks, visits to documents and sites

What happens when you can label billions of Web pages with reading difficulty metadata?

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

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

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**

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

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

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 ded-icate 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

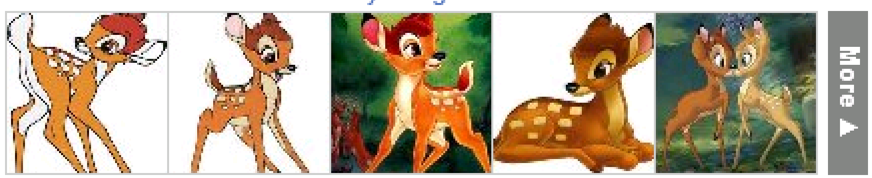
Cinderella
I bought **Cinderella**, and it is running in German only. I have a Mac. **Cinderella** does not run on my Computer, although I know I have Java 2 installed on it [usually ...
cinderella.de/tiki-view_faq.php?faqId=1

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

[ban

The SETI-Capable BAMBI Radio Telescope

By Bob Lash and Mike Fremont

Abstract

The design, construction, and initial observational results of a
4 GHz amateur radio telescope are described in this first report from
Project BAMBI (Bob And Mike's Big Investment). The system is now operating continuously. The planned extension of the BAMBI project to amateur SETI is also discussed.

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

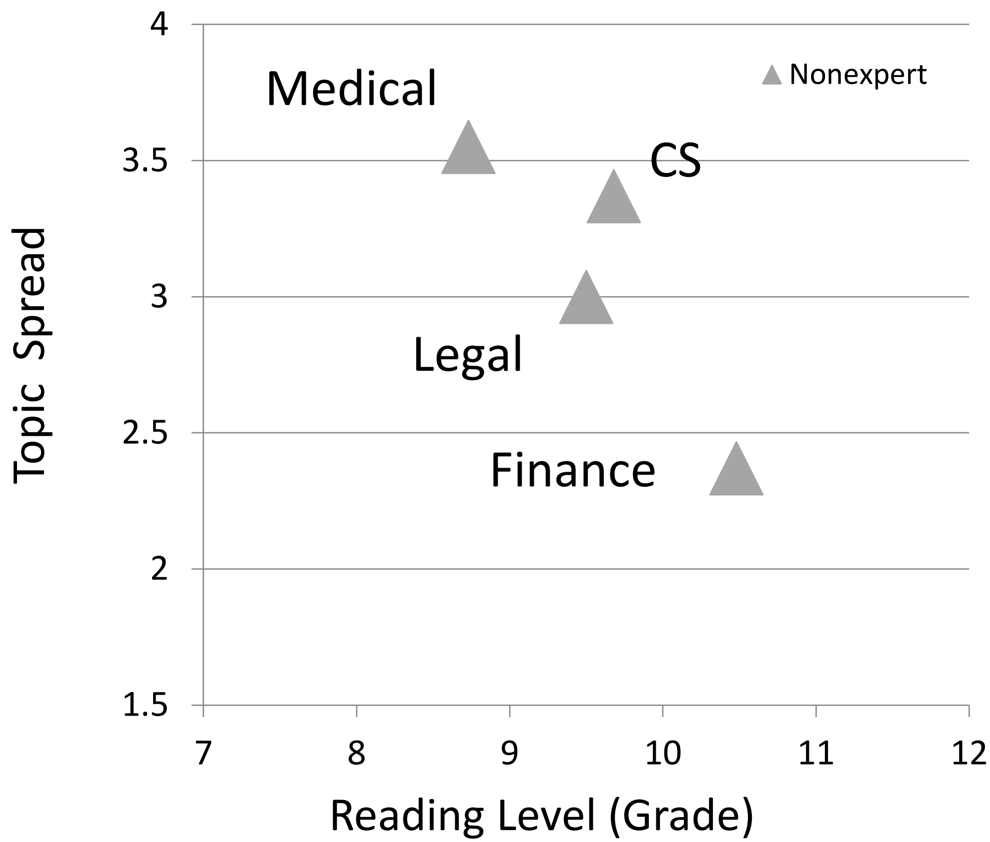
Adding reading level metadata to existing topic predictions helps model user and site expertise

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

- Four features:
  - Reading level
    1. Expected reading level *E*(*R*)
    2. Entropy *H*(*R*) of reading level distribution
  - Topic
    3. Top-*K* ODP category predictions
    4. Entropy *H*(*T*) of ODP category distribution
- Applications:
  - Classify expert *vs* non-expert sites and users
  - Better personalization features for search
  - Better click entropy prediction from topic/RL entropy
  - Spam/content farm detection
  - Detecting difficult tasks for users

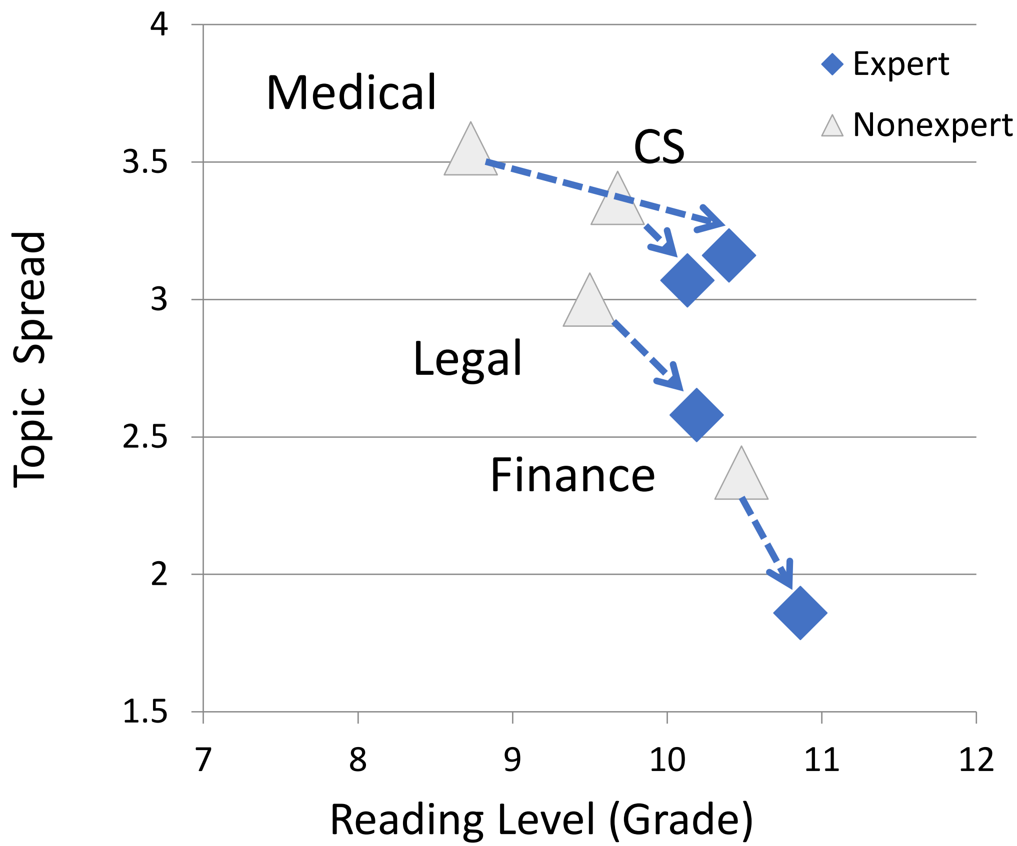
Reading level and topic entropy features can help separate expert from non-expert websites

[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]



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

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

From search sessions with unusually high d…
> 4 levels deviation fr…
user’s average

Highest association wi… st…tch reading	
Title …	L…
…sts	
…is	
plan	
pay	1.…
medical	1.69
learning	1.62

…ical tests
…lege entrance
…ncial aid
C… forms
Job s…rch

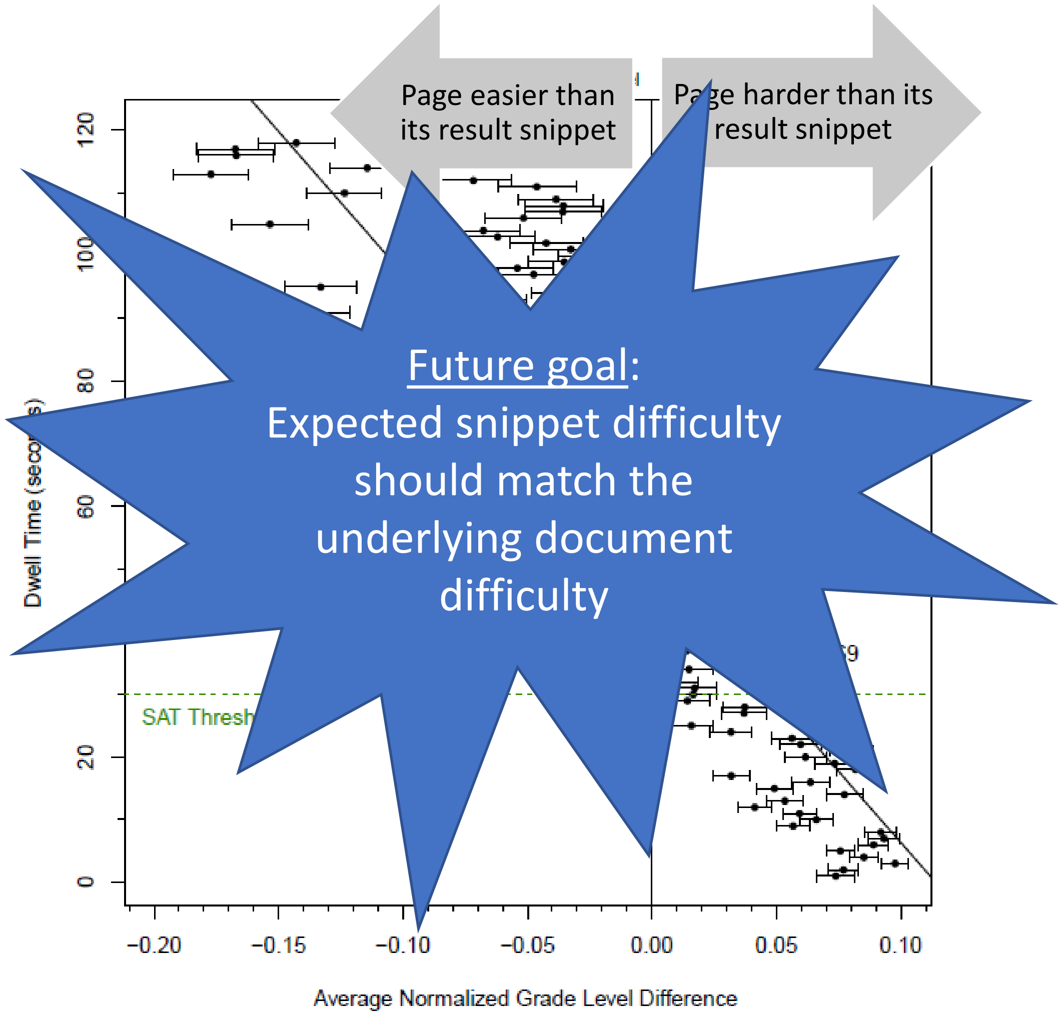
Future work:

1. Identify & predict stretch tasks
2. How/when to provide support?
3. Find helpful alternatives

Based on two months of Bing traffic.

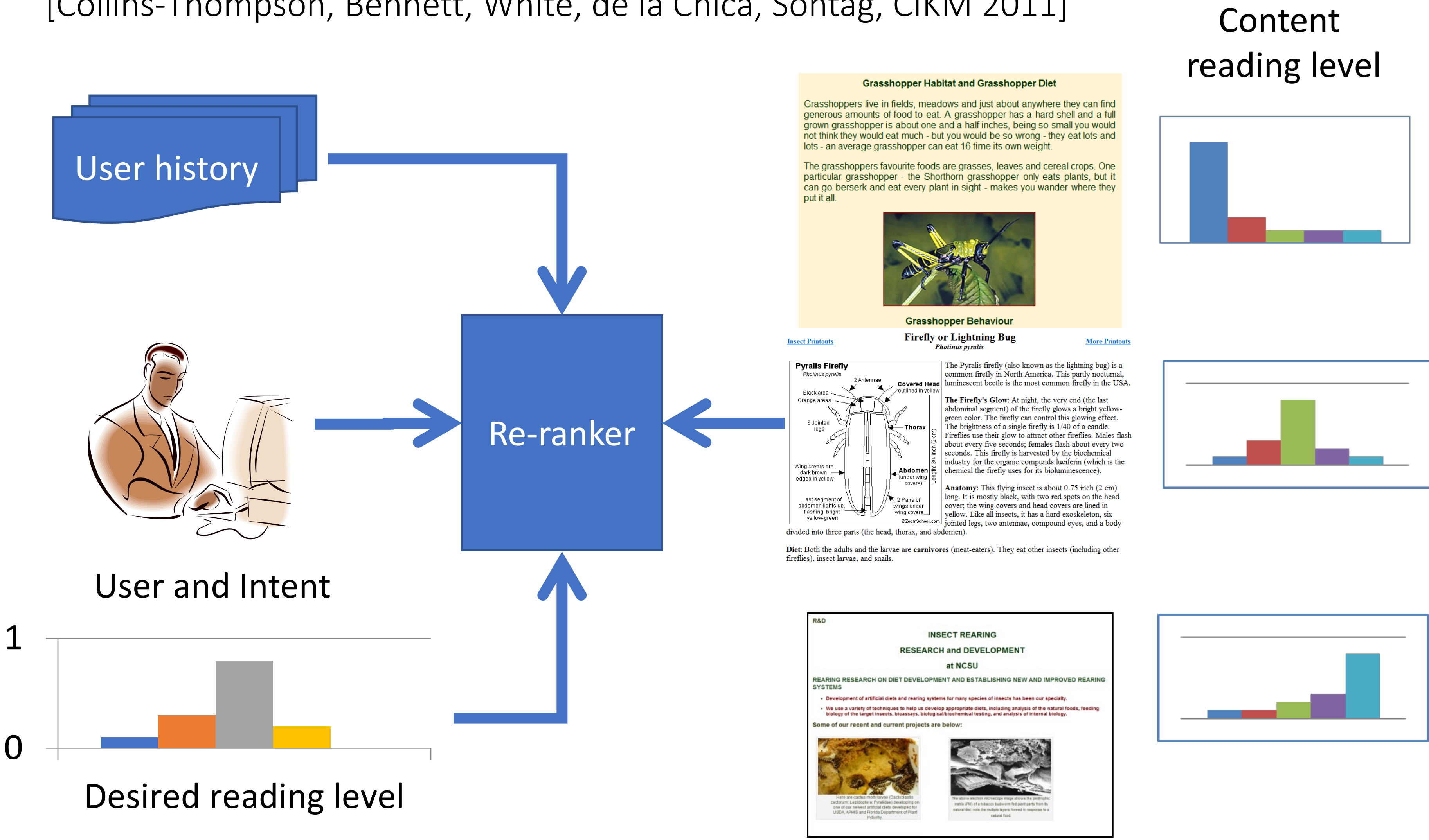
Adding reading level metadata improves click models for ranking.
Users can be misled by a mismatch between snippet difficulty & page difficulty
[Collins-Thompson, Bennett, White, de la Chica, Sontag, CIKM 2011]

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

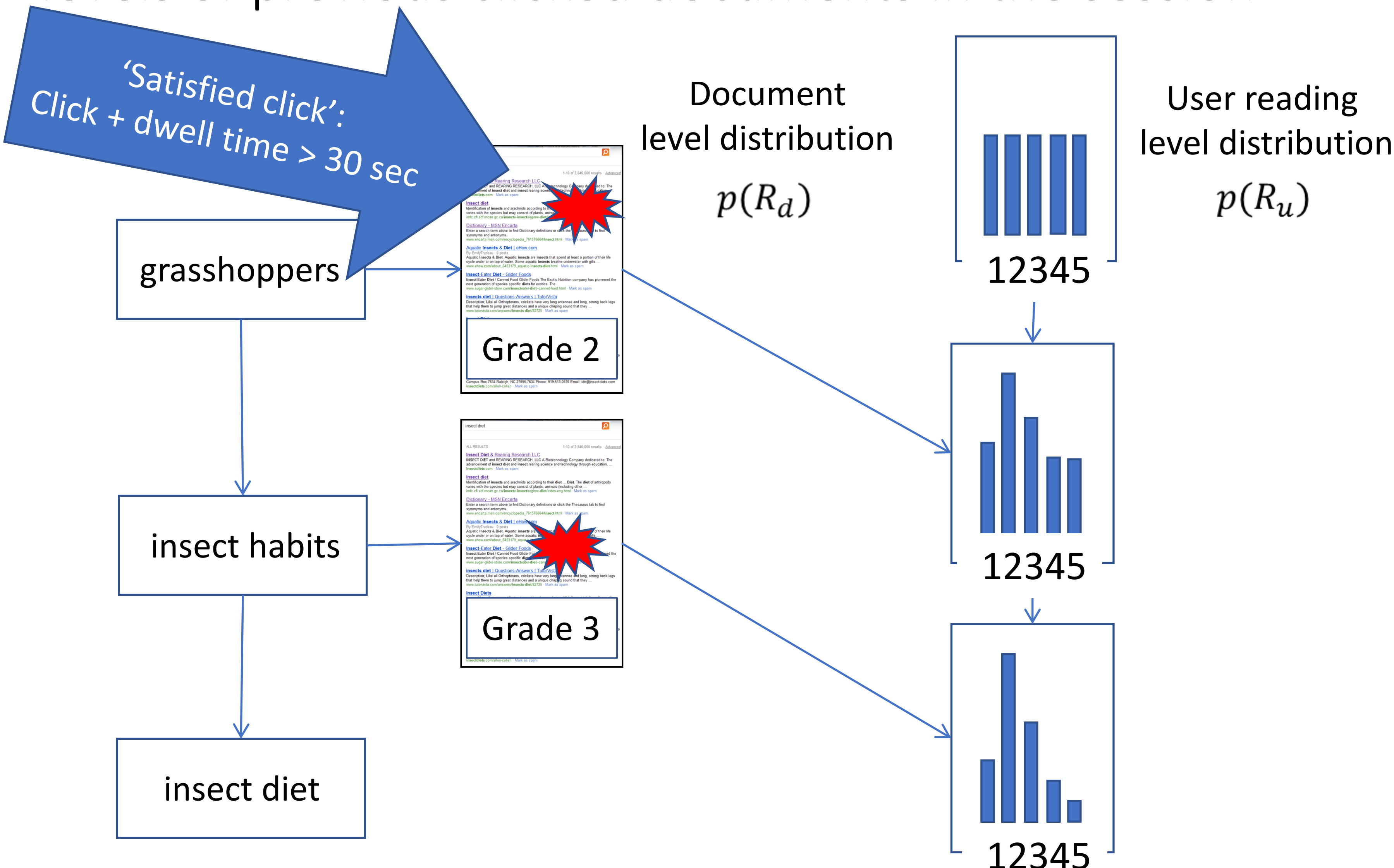


Personalizing Web search by reading level

[Collins-Thompson, Bennett, White, de la Chica, Sontag, CIKM 2011]



A simple generative user model combines reading levels of previous clicked documents in the session



Which features could help personalize?

- Content
  - Page reading level. (query-agnostic)
  - Result snippet reading level. (query-dependent)
- Query
  - Length in words, characters.
  - Query term reading level.
- User Session
  - All queries since last >30 min of inactivity.
  - Pages with satisfied clicks.

    <u>Assumption</u>: user likes results <= their desired level
- Interaction features
  - e.g. Snippet-Page difference in reading level.
- Variance of the above features.

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

Query subset	Num. queries	% Total	Gain

Gain from general relevance features: + 0.5
Gain from personalization features: + 0.7

Po[illegible]click relative to Bing production baseline.

- Gain for all queries; varied with query subset
  - Any gain ≥ 1.0 over production is notable.
  - Science queries benefited most.
- Net +1.6% of <u>all</u> queries improved at least one rank position in satisfied click
  - Large rank changes (> 5 positions) more than 70% likely to result in a win

Using reading level and word acquisition models to optimize search ranking for human learning

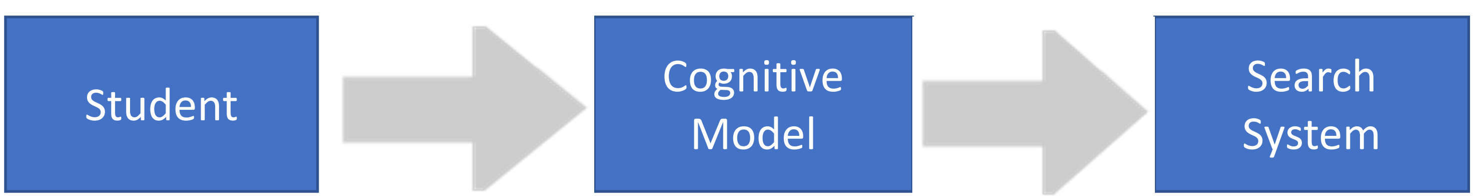
[Syed & Collins-Thompson, SIGIR 2017]

Key idea:

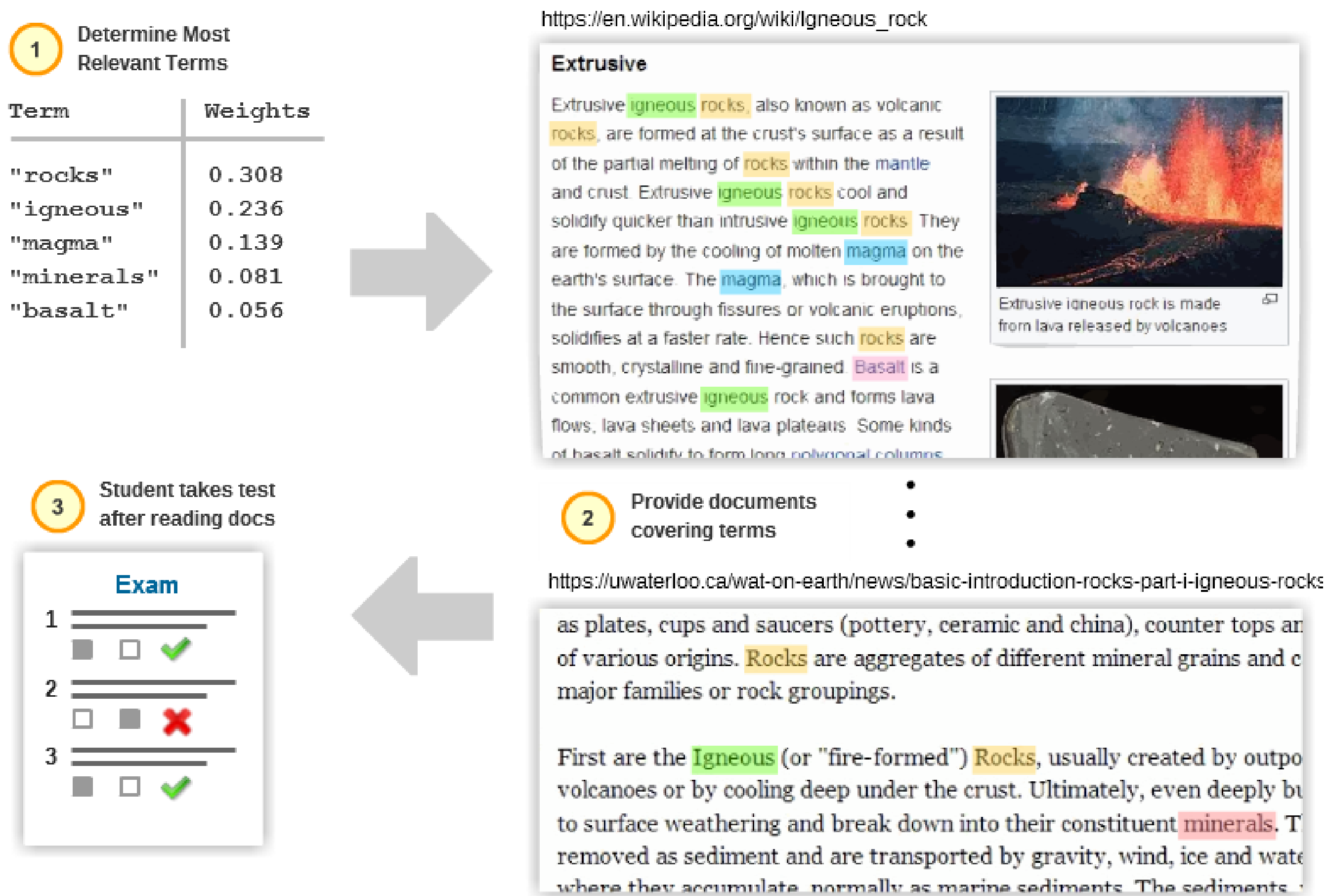
Optimize retrieval <u>through the cognitive lens of the user</u>

- Define representation of user's knowledge state
- Estimate prior and goal knowledge states
- Specify a cognitive model of how information affects the user's knowledge state.

i.e. how the user learns from the presented information

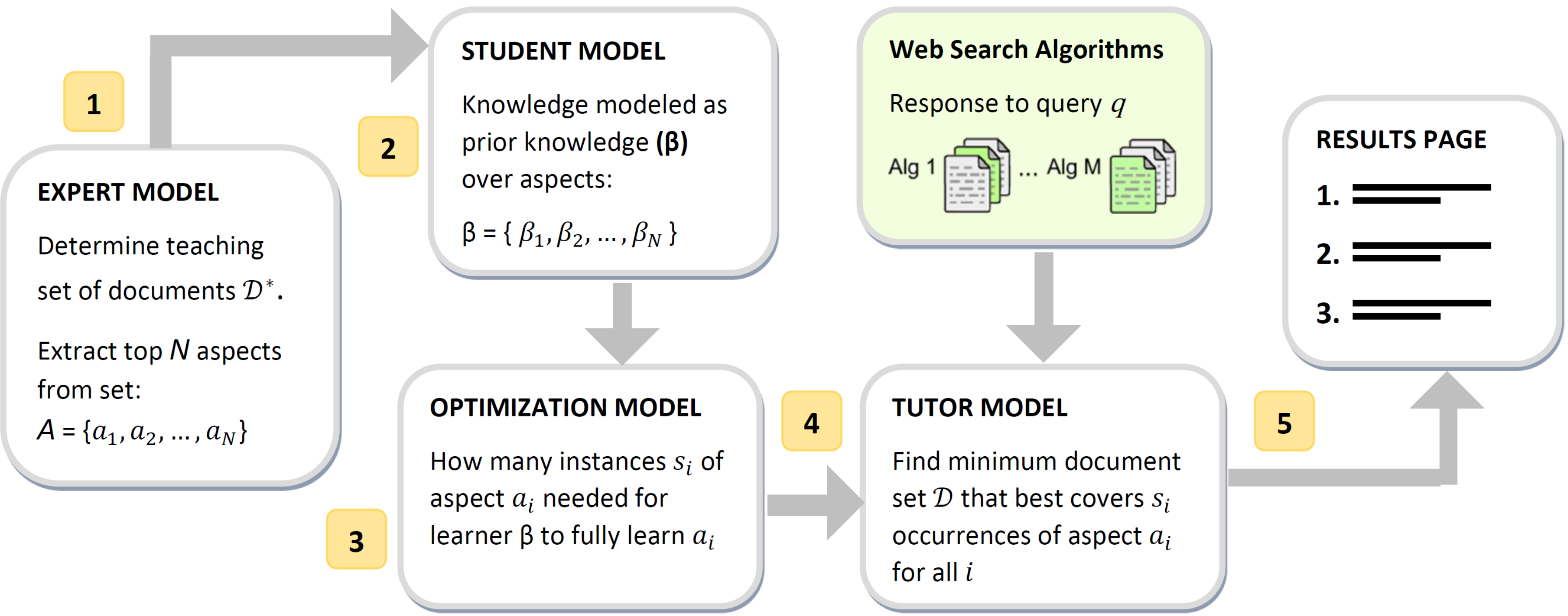


Vocabulary learning from context for the topic “igneous rock”

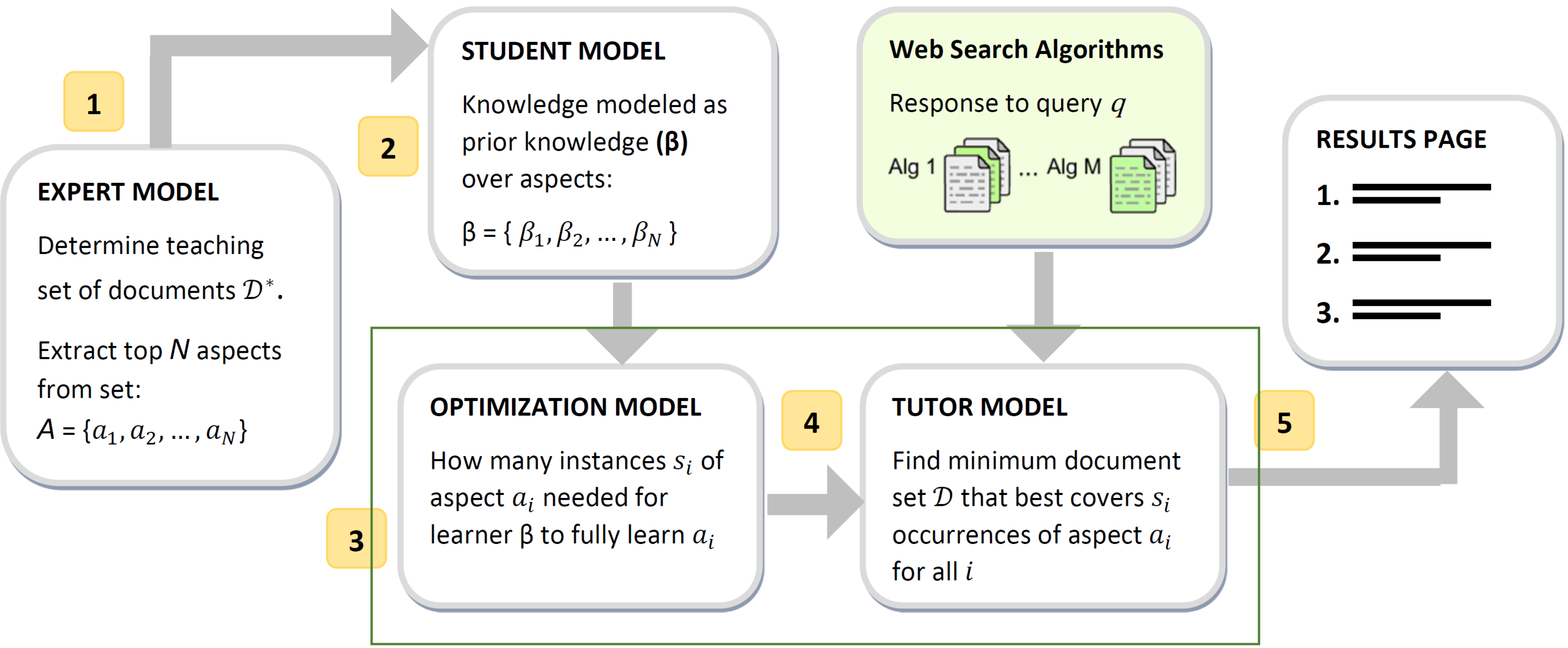


[Syed and Collins-Thompson, 2017]

Goal: Search algorithms that support effective vocabulary learning from context



Our SIGIR 2017 paper addressed the personalized retrieval problem in steps 3 and 4.



Optimally connecting users with information for personalized learning

What should a retrieval objective for learning look like? Find information that…

1. ***Advances an individual user's <u>learning progress</u> toward a specific goal.***

2. ***Minimizes or reduces <u>effort</u> toward that goal.***

Overall retrieval optimization problem: find an optimal set of documents *D*

$$argmax_D \; U(D) = H(D) - \lambda \cdot E(D)$$

Learning utility

Learning effectiveness

Learning effort

Searching efficiently
for optimal set *D* can be hard…

Search for an approximately optimal document set *D* into two steps

(a) How many total exposures S_k *for each word* are needed for each item *k* to efficiently maximize the learning outcome?

(b) Find a 'good' set of documents that exposes the user to these optimal S_k per-word exposures in context.

(c) User effort modeled by length and reading difficulty of the material.

Personalized ranking gave significantly better learning outcomes than generic search (and best overall)

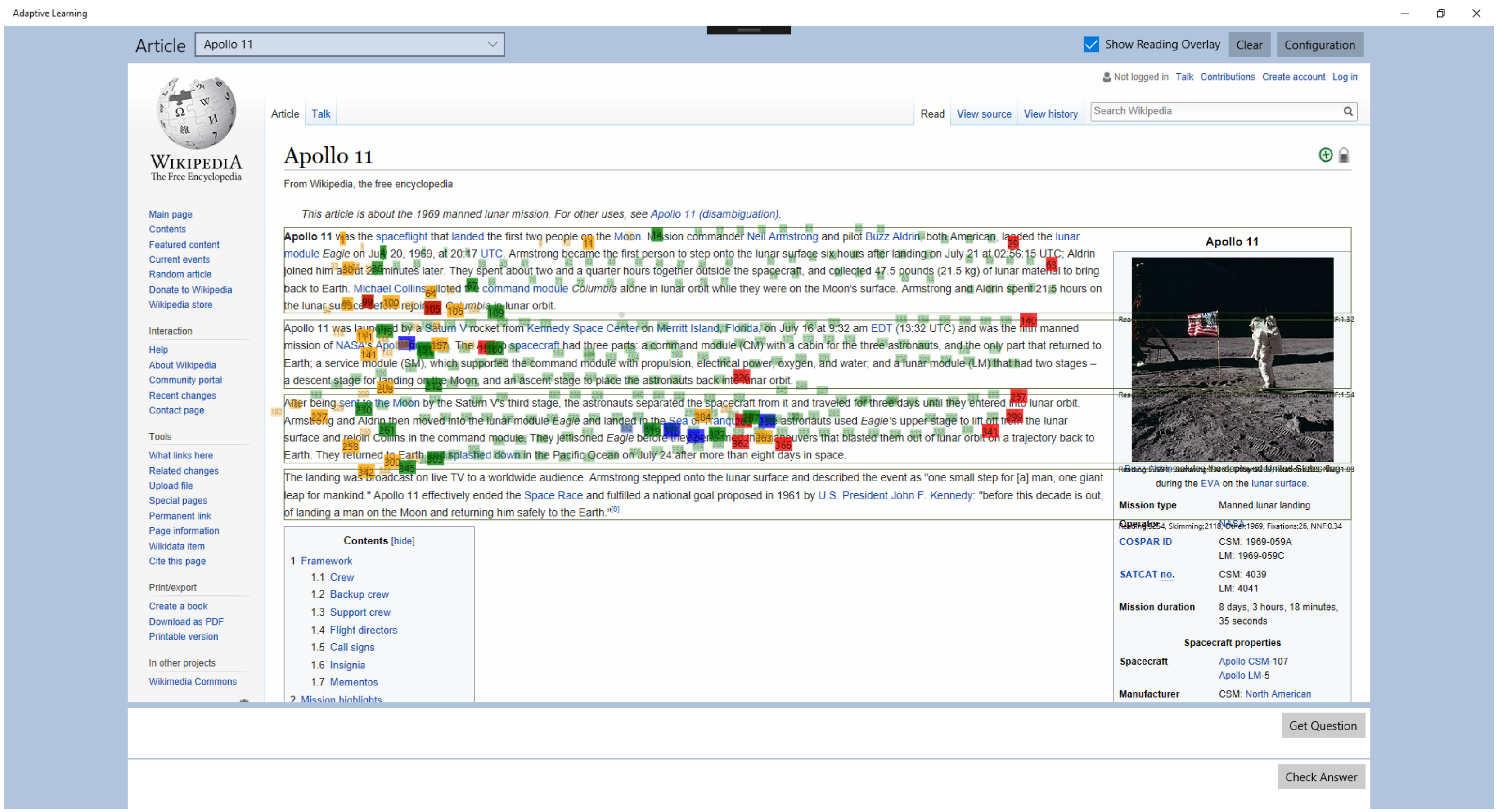
Measure	Web (n=290)	Non-Personalized (n=290)	Personalized (n=283)	p-value
Absolute Learning Gains	1.72	1.83	**1.98**	$p=.046$
Realized Potential Learning	0.38	0.43	**0.47**	$p=.008$
Learning Gains/ 1000 Words	0.11	0.25	**0.35**	$p<.001$

Type of Test: Kruskal Wallis H Test - Omnibus

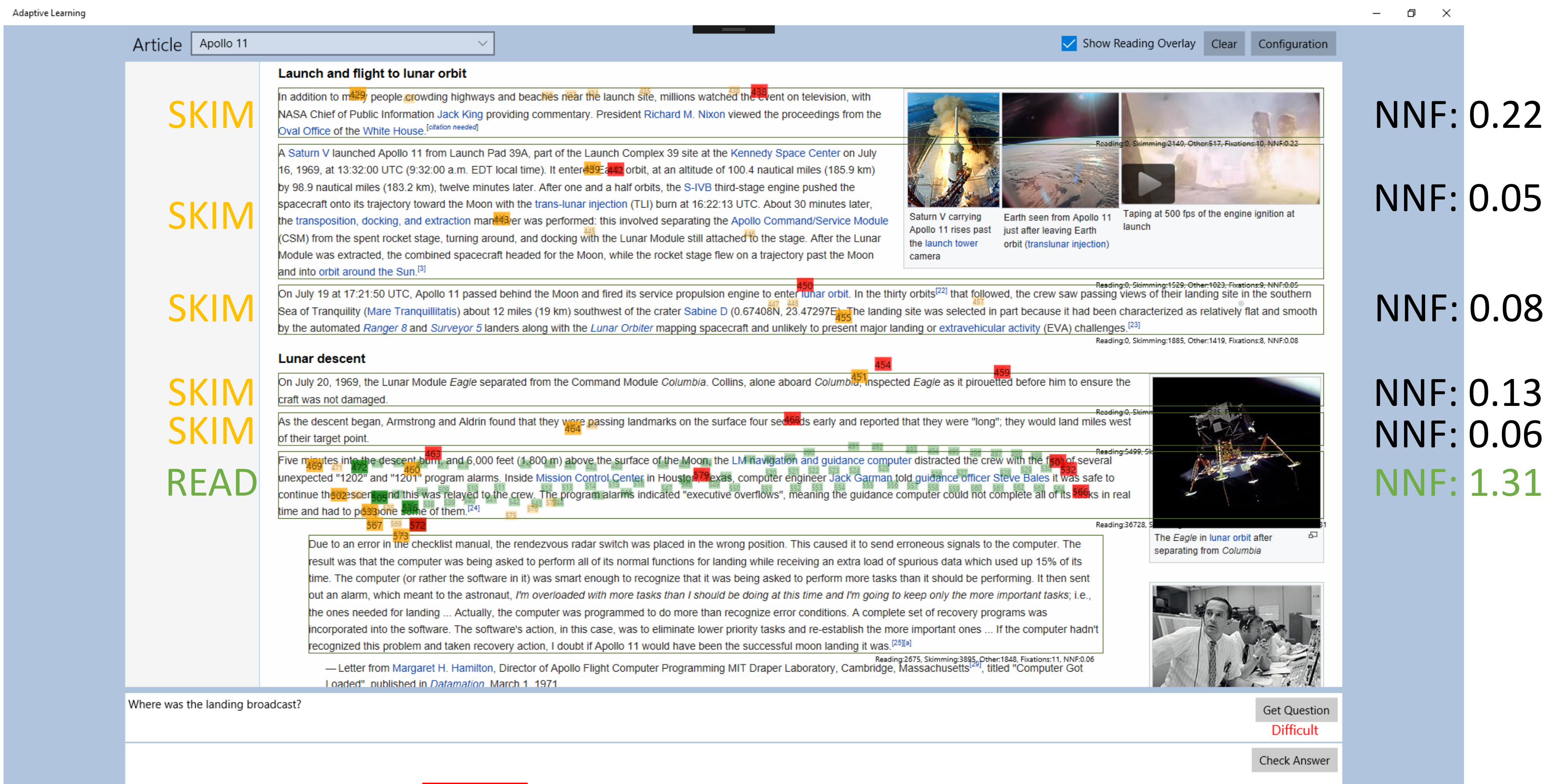
Personalization helps long-term learning, confirmed also in:
CHIIR 2018: Exploring Document Retrieval Features Associated with Improved Short- and Long-term Vocabulary Learning Outcomes
[Syed and Collins-Thompson, 2018]

Adaptive Learning for Reading Comprehension: Gaze classifier tracks reading fixations, generates adjunct questions for the reader

[Syed, Collins-Thompson, Bennett, Teng, Williams, Tay, Iqbal. WWW 2020]

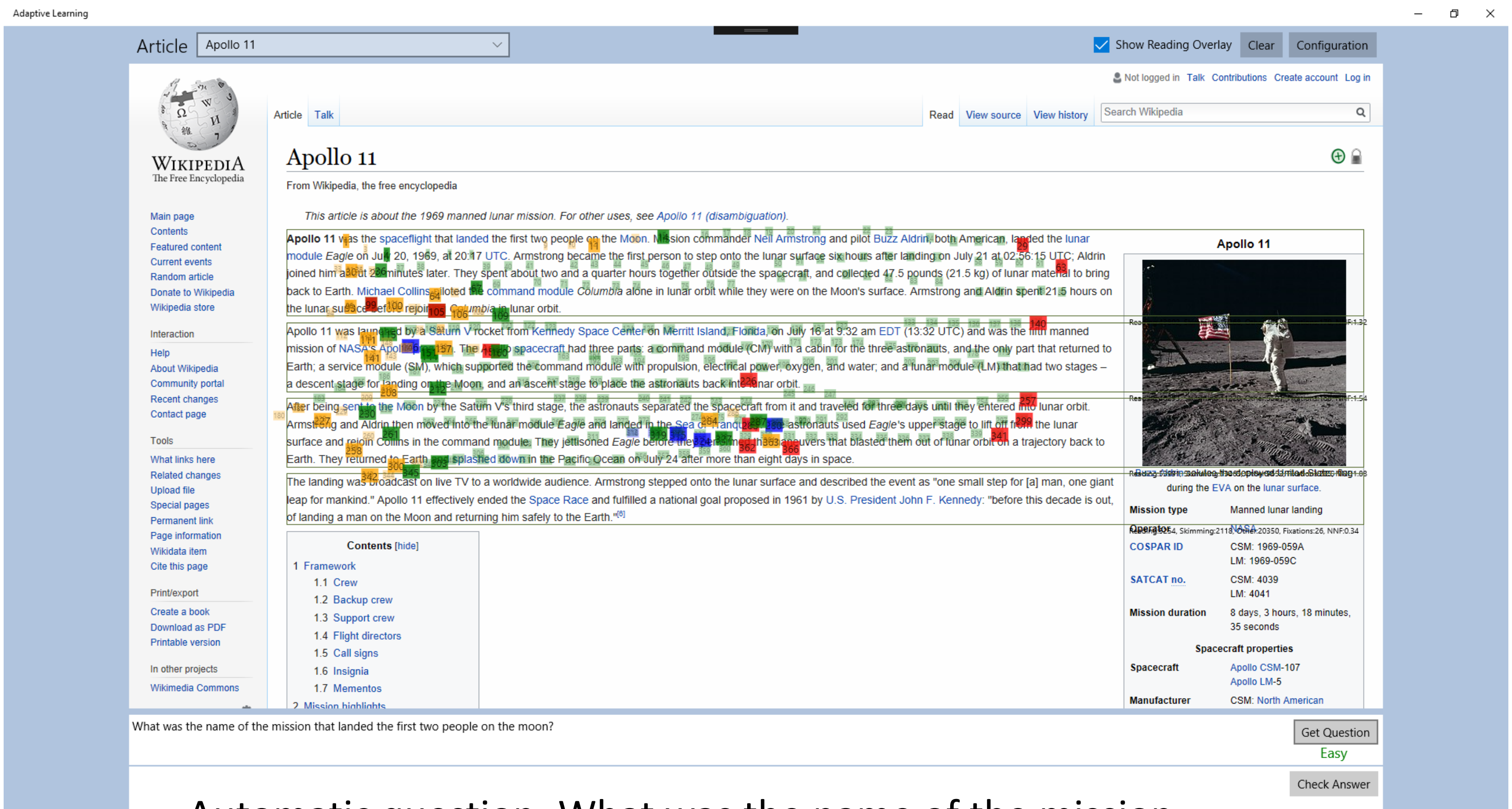


Attention is approximately quantified by **N**ormalized **N**umber of **F**ixations (per word) in content window: Fixation types: skimming, reading, regression

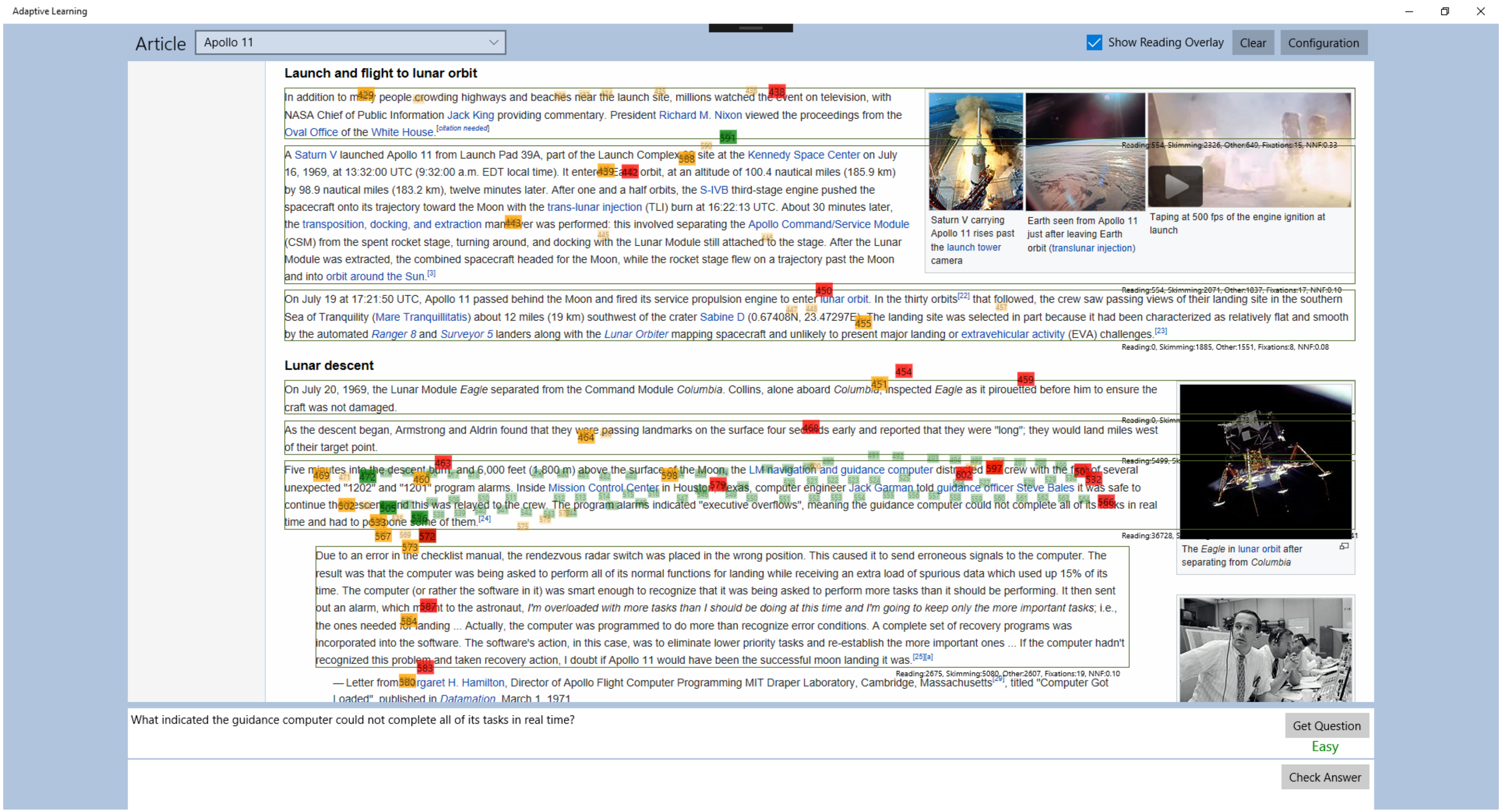


Regression fixations indicate material that a user had to go back and re-read

Easy questions are generated from content read more closely

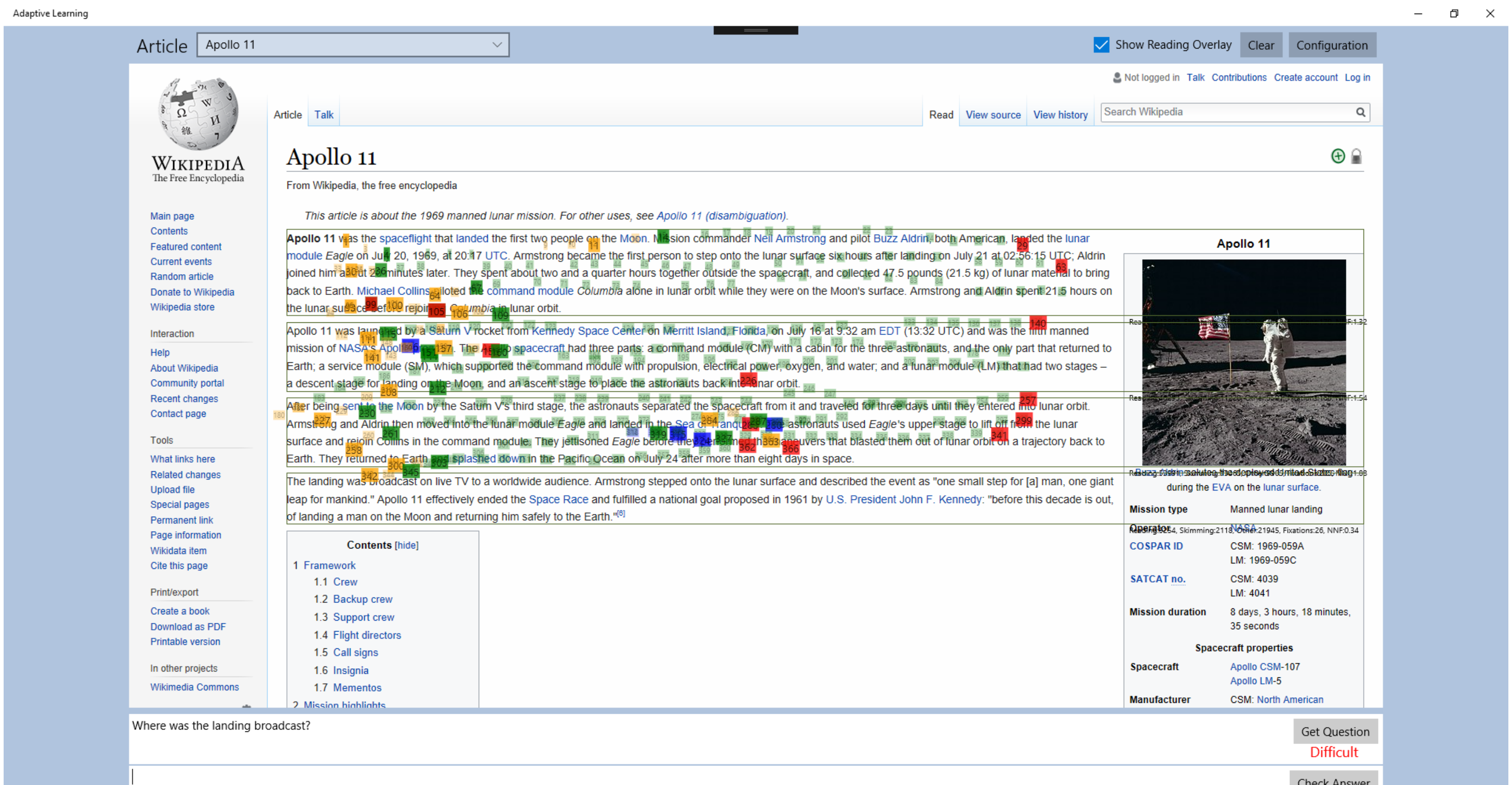


Automatic question: What was the name of the mission…



Automatic question: What indicated the guidance computer could not complete all of its tasks in real time?

Difficult questions are generated from content that was skimmed



Automatic question: Where was the landing broadcast?

Most significant learning effects were detected in the long-term condition, especially for low-knowledge learners
[*Syed, Collins-Thompson, Bennett, Teng, Williams, Tay, Iqbal. WWW 2020*]

Result	**Short-term learning**	**Long-term retention**
Adjunct Questions improved grades better than $\mathbf{Q_{None}}$ (Section 5.2)	No	Yes (for low-knowledge learners)
$\mathbf{Q_{Auto}}$ performed comparable to $\mathbf{Q_{Human}}$ (Section 5.3)	Yes	Yes
Synthesis question affected grades (Section 5.4)	No	No
Focus-based question selection improved grades (Section 5.5)	No	Yes
Gaze behavior was different for those who would answer questions correctly (Section 6.4)	Yes (for low-knowledge learners)	Yes (for low-knowledge learners)

- Key takeaway: always have a delayed post-test!

Selected next steps

1. Domain-specific, personalized measures of content difficulty.
2. Modeling <u>desirable</u> difficulty in content.
3. Search and recommendation for educational videos.
   a. Multi-modal documents: transcript, slides, lecture images, audio, supporting materials…
   b. Local difficulty and complexity in lecture segments
   c. Lots of business applications also: connect with OCR
4. Deep learning models of contextual informativeness.
   a. Finding supportive text for learning new concepts.
   b. Curriculum learning for *machine* readers.
   c. <u>Explainable</u> models.

Our basic statistical readability models are available via REST API for non-commercial research use:

api.dscovar.org

Long-term: Modeling content difficulty for better support of human learning

1. Develop richer representations of users and content to support learning.
2. Integrate rich models of human learning into search and recommender algorithm objectives, features, and evaluation.
3. Continue developing reliable automated methods for explicit and implicit assessment of difficulty and learning during interaction.
4. Aim for robust long-term retention, not just short-term learning.

Contact: Kevyn Collins-Thompson

kevynct@umich.edu
Readability API: api.dscovar.org
Homepage: www.umich.edu/~kevynct

This research supported in part by:

This research supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A140647 to the University of Michigan. The opinions expressed are those of the author and do not represent views of the Institute or the U.S. Department of Education.