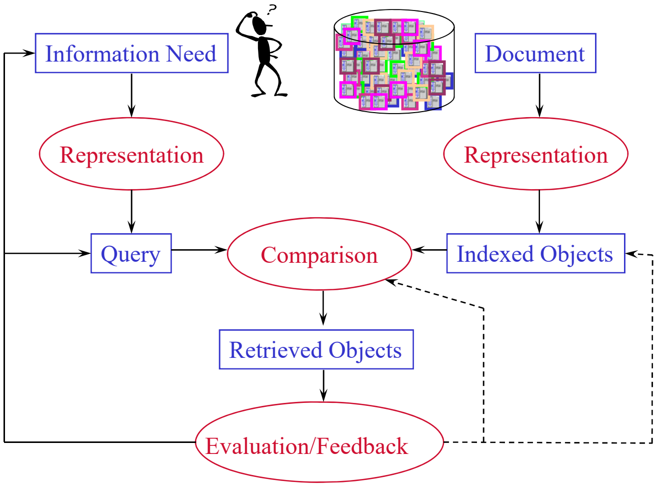
11-642:
Search Engines

Information Needs and Queries

Jamie Callan
Carnegie Mellon University
callan@cs.cmu.edu

Overview of Information Retrieval Processes


Information Need

Document

Representation

Representation

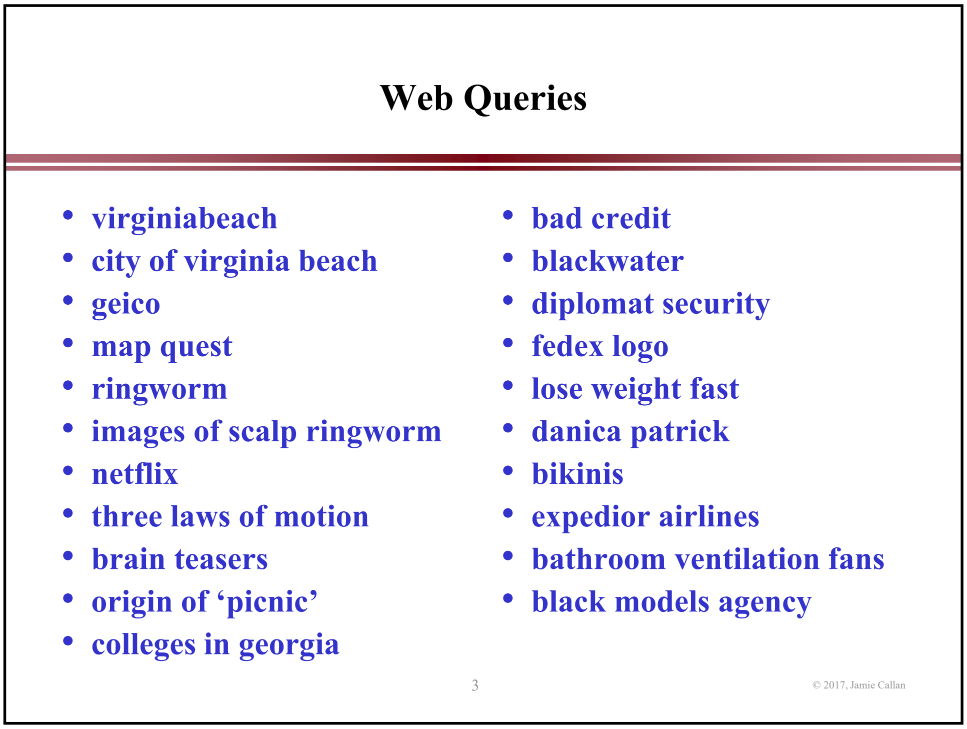
Query

Comparison

Indexed Objects

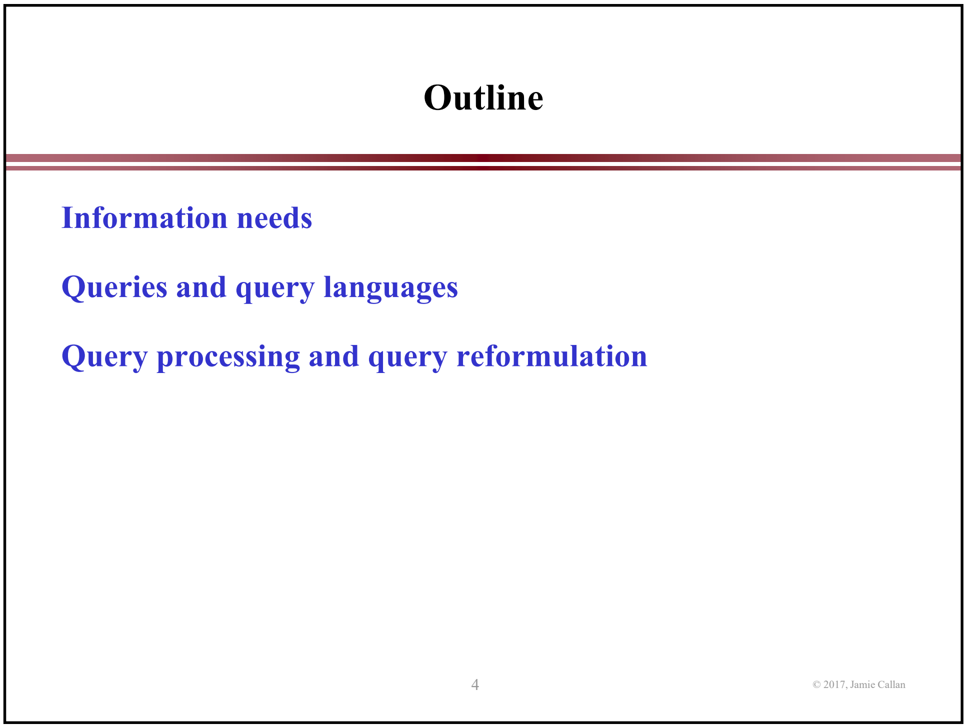
Retrieved Objects

Evaluation/Feedback



Web Queries

- virginiabeach
- city of virginia beach
- geico
- map quest
- ringworm
- images of scalp ringworm
- netflix
- three laws of motion
- brain teasers
- origin of ‘picnic’
- colleges in georgia
- bad credit
- blackwater
- diplomat security
- fedex logo
- lose weight fast
- danica patrick
- bikinis
- expedior airlines
- bathroom ventilation fans
- black models agency



Outline

Information needs

Queries and query languages

Query processing and query reformulation

Information Needs

A person begins a search with an <u>information need</u> in mind

The information need is implicit and unknown

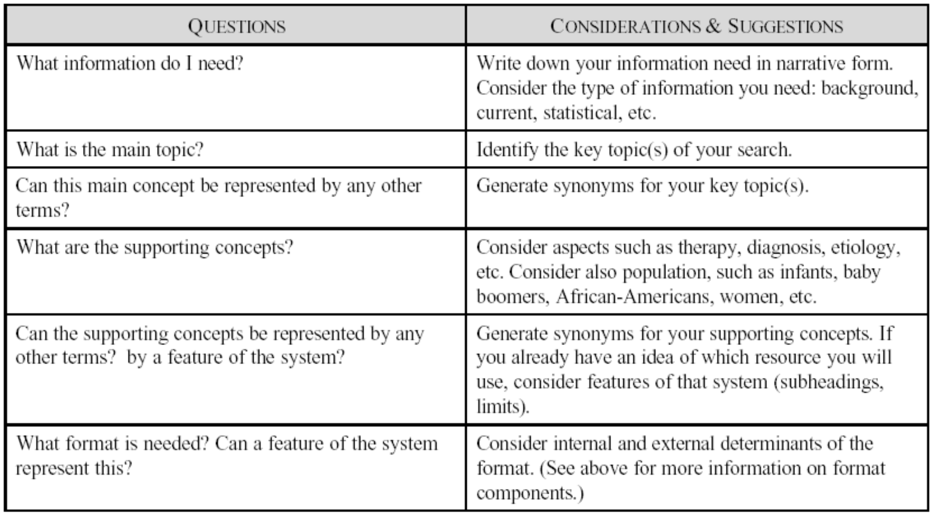
- The query describes the information need

  …but it may not be an <u>accurate</u> description

Often people don't describe their information needs well

- Librarians are trained to elicit information needs
- Much of what is known about this topic is from Library Science
  - How well does this information apply to the web?

Eliciting Information Needs



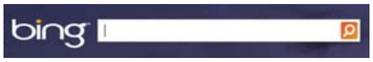
QUESTIONS	CONSIDERATIONS & SUGGESTIONS
What information do I need?	Write down your information need in narrative form. Consider the type of information you need: background, current, statistical, etc.
What is the main topic?	Identify the key topic(s) of your search.
Can this main concept be represented by any other terms?	Generate synonyms for your key topic(s).
What are the supporting concepts?	Consider aspects such as therapy, diagnosis, etiology, etc. Consider also population, such as infants, baby boomers, African-Americans, women, etc.
Can the supporting concepts be represented by any other terms? by a feature of the system?	Generate synonyms for your supporting concepts. If you already have an idea of which resource you will use, consider features of that system (subheadings, limits).
What format is needed? Can a feature of the system represent this?	Consider internal and external determinants of the format. (See above for more information on format components.)

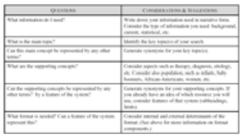
© 1999-2002 the Raymon H. Mulford Library, Medical College of Ohio.

Information Needs

Search engines also elicit information needs

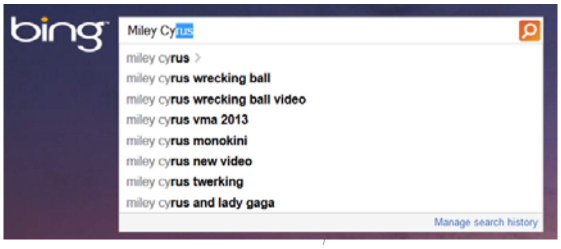
- How do they compare to elicitation by librarians?
- Initial elicitation





Questions	Considerations & Suggestions
What information do I need?	Write down your information need in narrative form. Consider the type of information you need: background, current, statistical, etc.
What is the main topic?	Identify the key topic(s) of your search.
Can this main concept be represented by any other terms?	Generate synonyms for your key topic(s).
What are the supporting concepts?	Consider aspects such as therapy, diagnosis, etiology, etc. Consider also population, such as infants, baby boomers, African-Americans, women, etc.
Can the supporting concepts be represented by any other terms? By a feature of the system?	Generate synonyms for your supporting concepts. If you already have an idea of which resource you will use, consider features of that system (subheadings, limits).
What format is needed? Can a feature of the system represent this?	Consider internal and external determinants of the format. (See above for more information on format components.)

- Subsequent elicitation



Miley Cyrus

miley cyrus
miley cyrus wrecking ball
miley cyrus wrecking ball video
miley cyrus vma 2013
miley cyrus monokini
miley cyrus new video
miley cyrus twerking
miley cyrus and lady gaga

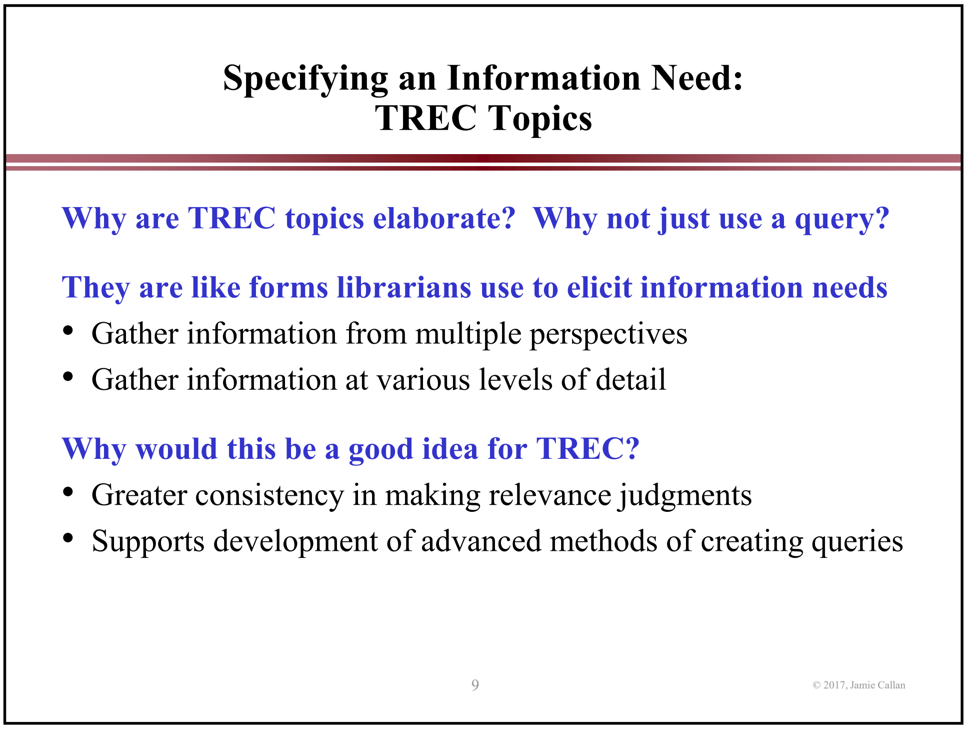
Specifying an Information Need: TREC Blog Track Topic 1105

Query: parenting

Description: I am looking for blogs that provide advice, counseling, and information on parenting.

Facet: personal

Narrative: Relevant blogs include those from parents, grandparents, or others involved in parenting, raising, or caring for children. Blogs can include those provided by health care providers if the focus is on children. Blogs that serve primarily as links to other sites, or that of themselves, market products related to children and their caregivers, are not relevant.



Specifying an Information Need: TREC Topics

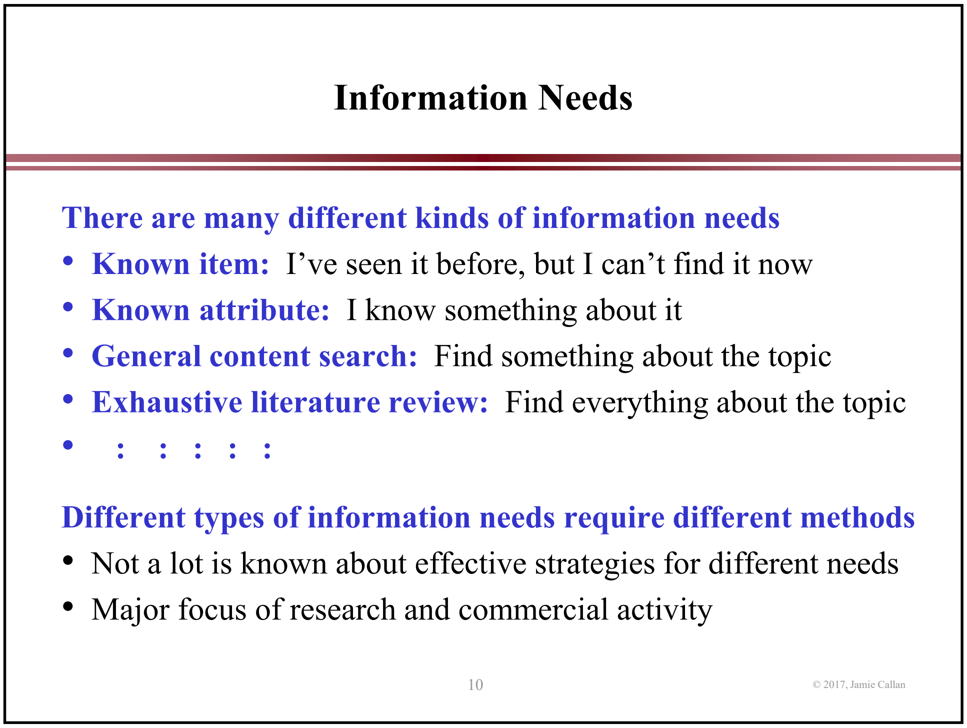
Why are TREC topics elaborate? Why not just use a query?

They are like forms librarians use to elicit information needs

- Gather information from multiple perspectives
- Gather information at various levels of detail

Why would this be a good idea for TREC?

- Greater consistency in making relevance judgments
- Supports development of advanced methods of creating queries



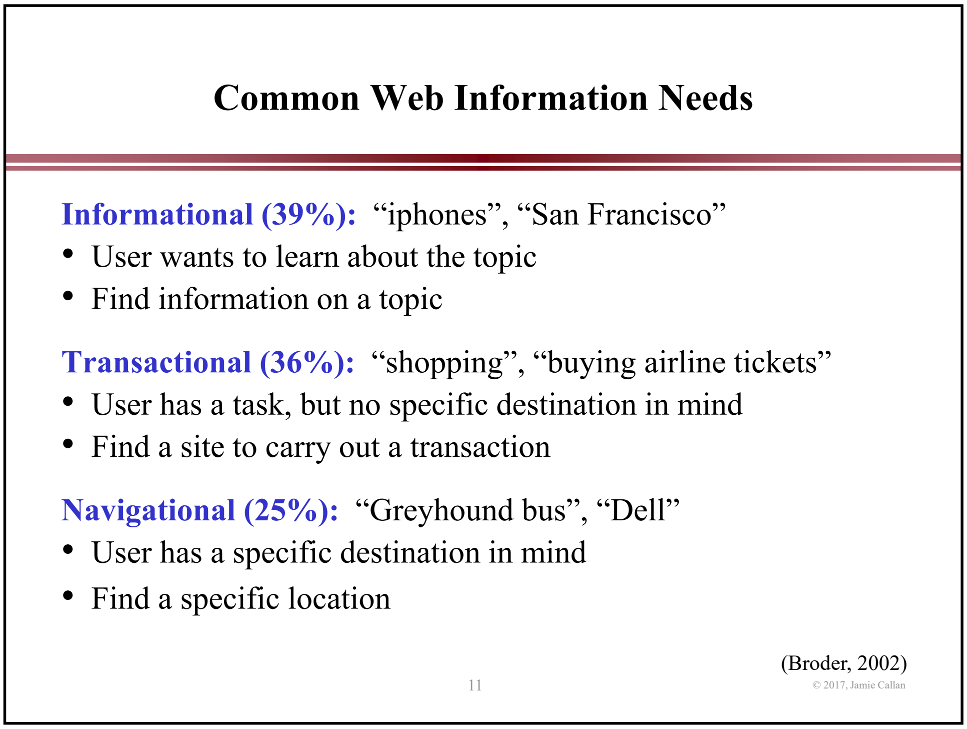
Information Needs

There are many different kinds of information needs

- **Known item:** I've seen it before, but I can't find it now
- **Known attribute:** I know something about it
- **General content search:** Find something about the topic
- **Exhaustive literature review:** Find everything about the topic
- : : : : :

Different types of information needs require different methods

- Not a lot is known about effective strategies for different needs
- Major focus of research and commercial activity



Common Web Information Needs

Informational (39%): “iphones”, “San Francisco”

- User wants to learn about the topic
- Find information on a topic

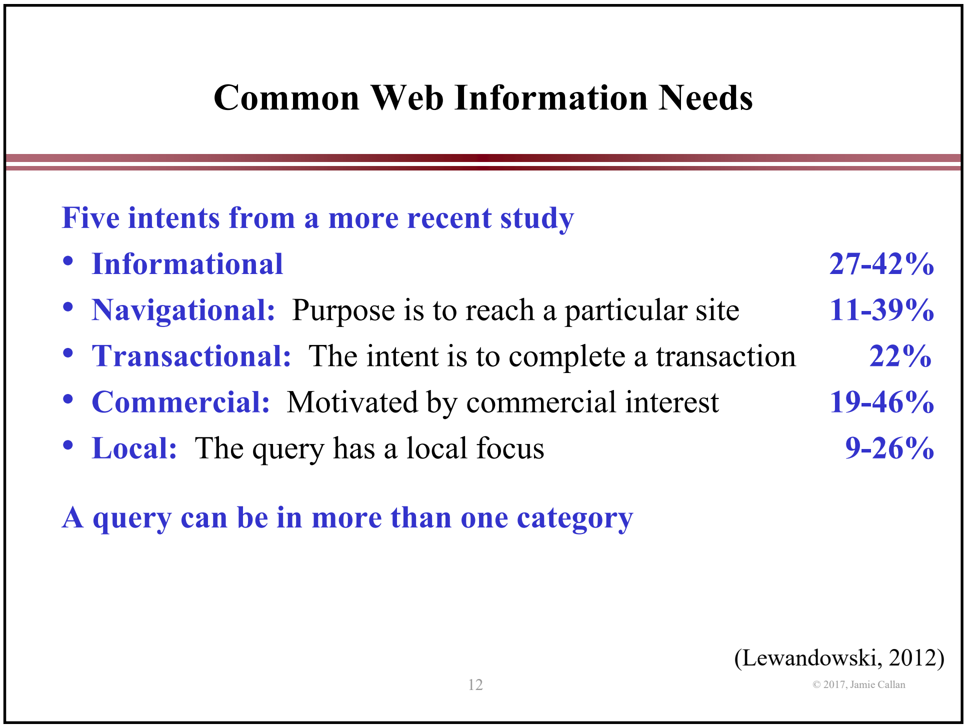
Transactional (36%): “shopping”, “buying airline tickets”

- User has a task, but no specific destination in mind
- Find a site to carry out a transaction

Navigational (25%): “Greyhound bus”, “Dell”

- User has a specific destination in mind
- Find a specific location

(Broder, 2002)



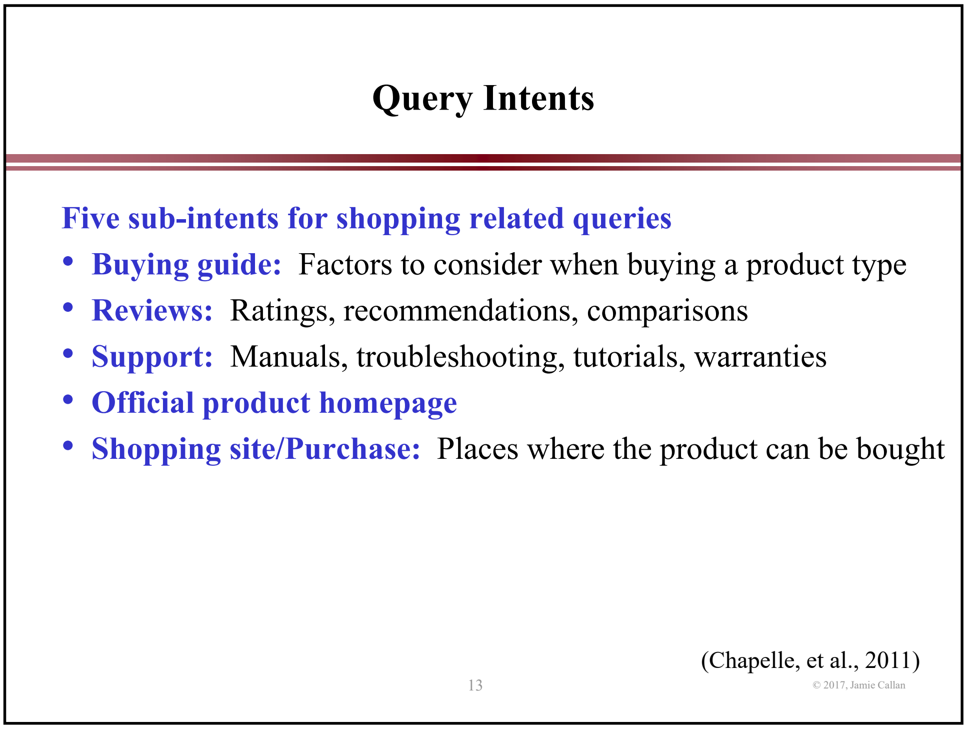
Common Web Information Needs

Five intents from a more recent study

- **Informational** **27-42%**
- **Navigational:** Purpose is to reach a particular site **11-39%**
- **Transactional:** The intent is to complete a transaction **22%**
- **Commercial:** Motivated by commercial interest **19-46%**
- **Local:** The query has a local focus **9-26%**

A query can be in more than one category

(Lewandowski, 2012)

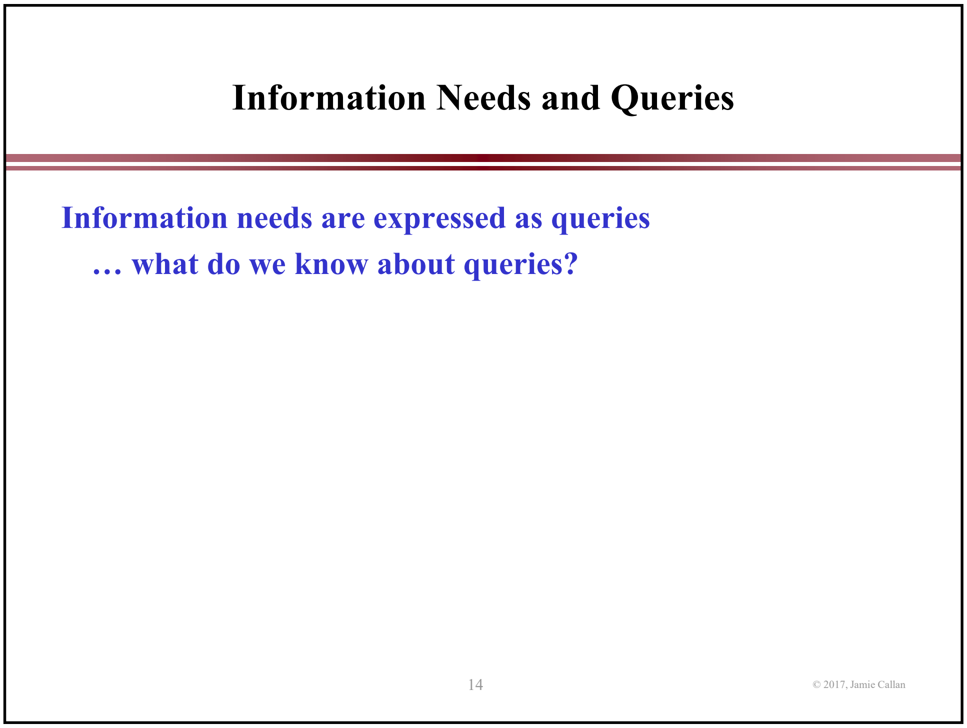


Query Intents

Five sub-intents for shopping related queries

- **Buying guide:** Factors to consider when buying a product type
- **Reviews:** Ratings, recommendations, comparisons
- **Support:** Manuals, troubleshooting, tutorials, warranties
- **Official product homepage**
- **Shopping site/Purchase:** Places where the product can be bought

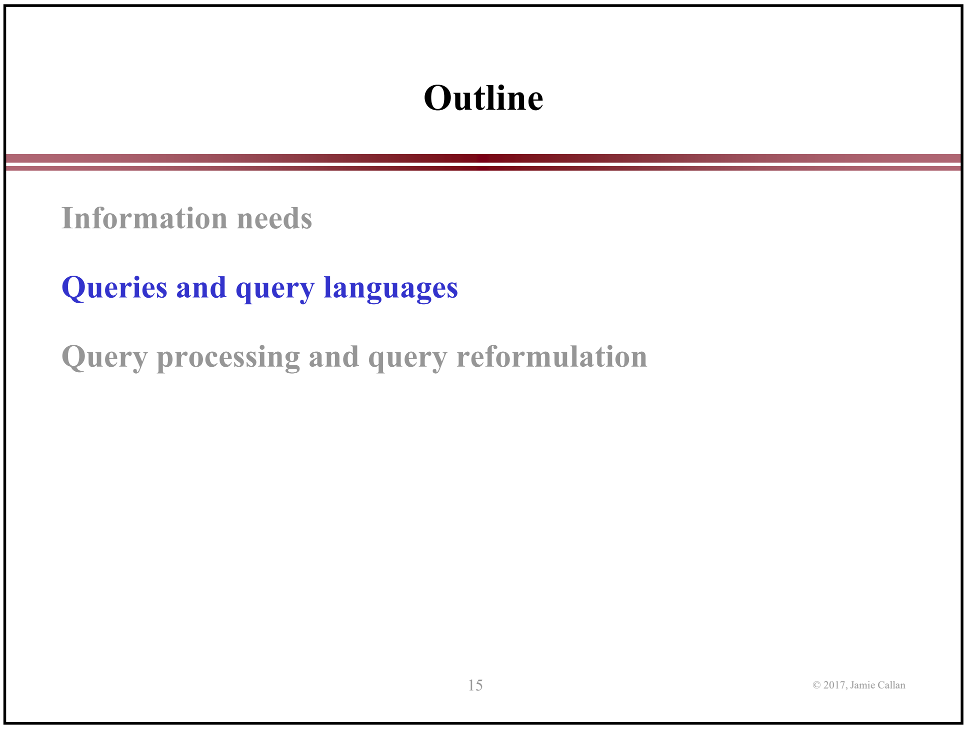
(Chapelle, et al., 2011)



Information Needs and Queries

Information needs are expressed as queries

… what do we know about queries?

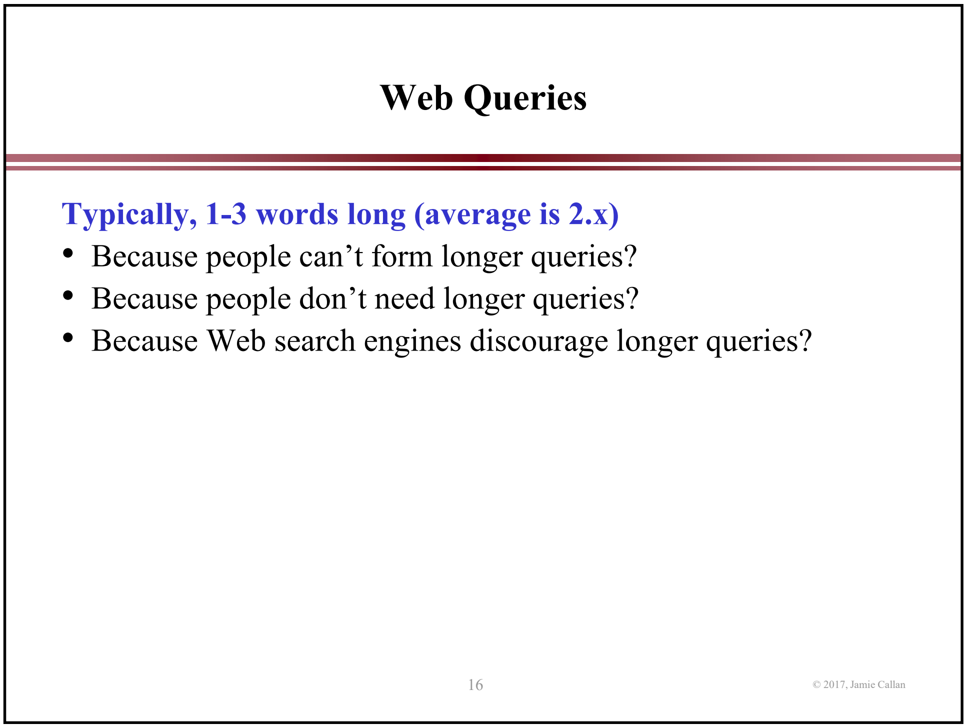


Outline

Information needs

Queries and query languages

Query processing and query reformulation



Web Queries

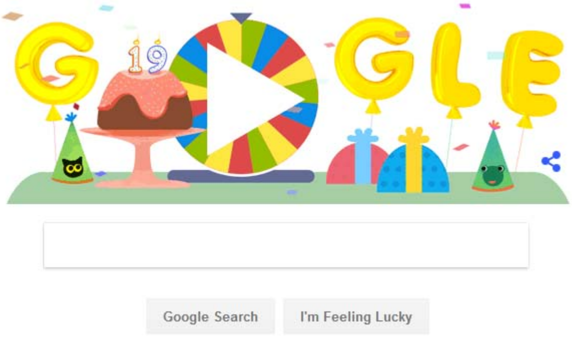
Typically, 1-3 words long (average is 2.x)

- Because people can’t form longer queries?
- Because people don’t need longer queries?
- Because Web search engines discourage longer queries?

Information Needs and Queries

The user interface plays a large role in how people express their information need

- A small box encourages short queries



Big picture (entertainment)

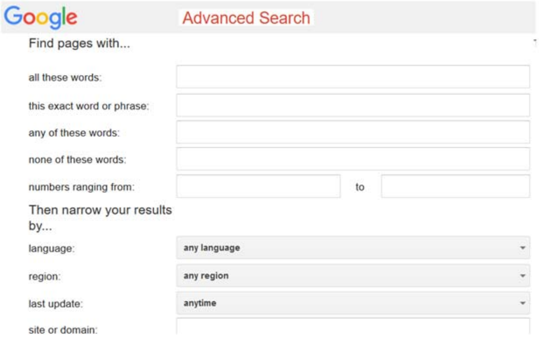
Small search box (the task)

1 /

Information Needs and Queries

The user interface plays a large role in how people express their information need

- A small box encourages short queries
- A form encourages more detail



Advanced Search

Find pages with...

all these words:

this exact word or phrase:

any of these words:

none of these words:

numbers ranging from: to

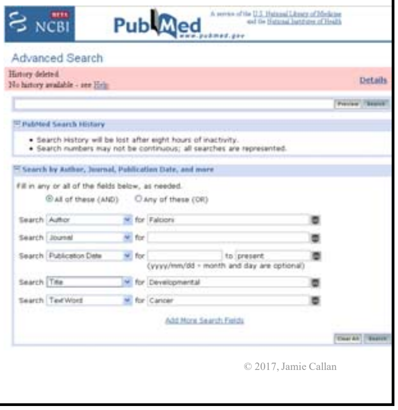
Then narrow your results by...

language: any language

region: any region

last update: anytime

site or domain:


NCBI
PubMed
www.pubmed.gov
A service of the U.S. National Library of Medicine and the National Institutes of Health
Details
Preview
Search
Add More Search Fields
Clear All
Search

Advanced Search

History deleted
No history available - see Help

PubMed Search History

- Search History will be lost after eight hours of inactivity.
- Search numbers may not be continuous; all searches are represented.

Search by Author, Journal, Publication Date, and more

Fill in any or all of the fields below, as needed.

● All of these (AND) ○ Any of these (OR)

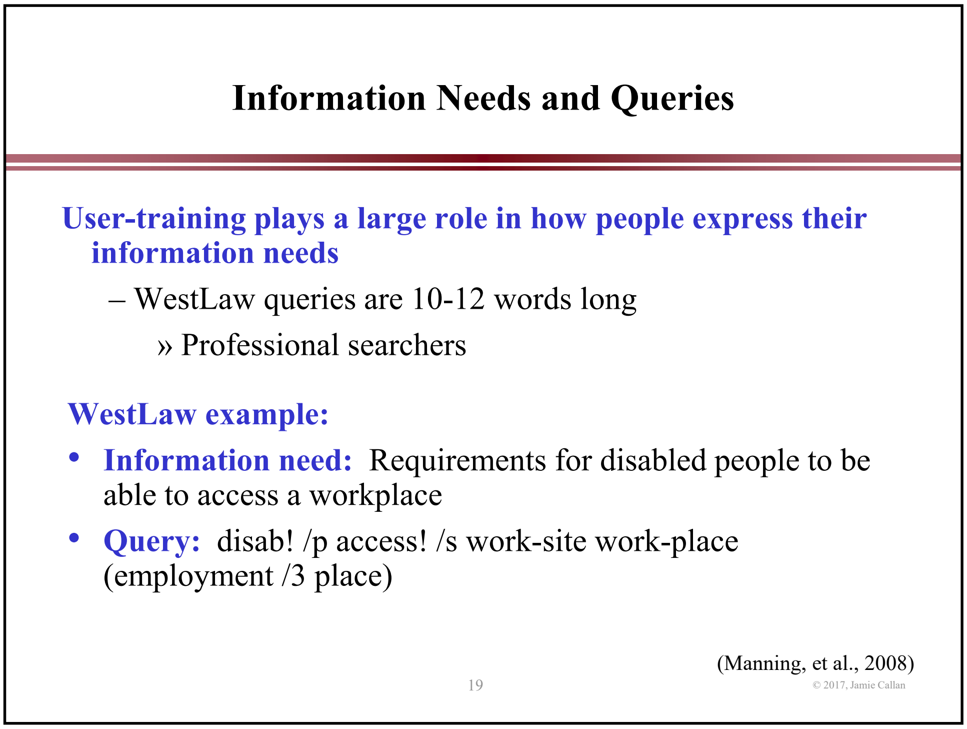
Search Author for Falcons

Search Journal for

Search Publication Date for to present
(yyyy/mm/dd - month and day are optional)

Search Title for Developmental

Search Text Word for Cancer



Information Needs and Queries

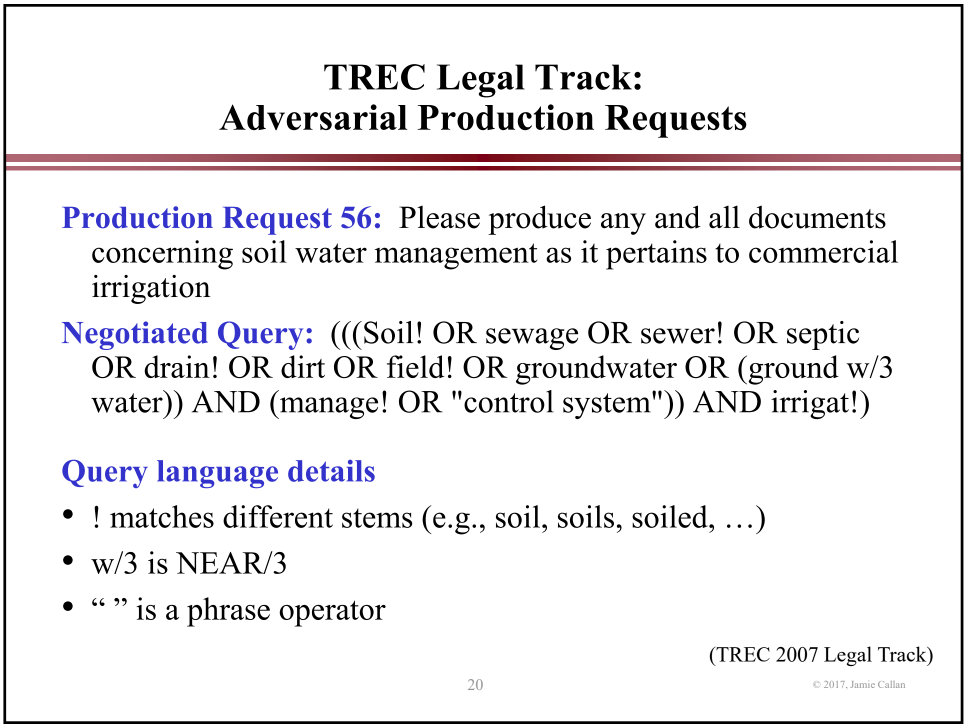
User-training plays a large role in how people express their information needs

- WestLaw queries are 10-12 words long
  - » Professional searchers

WestLaw example:

- **Information need:** Requirements for disabled people to be able to access a workplace
- **Query:** disab! /p access! /s work-site work-place (employment /3 place)

(Manning, et al., 2008)



TREC Legal Track:
Adversarial Production Requests

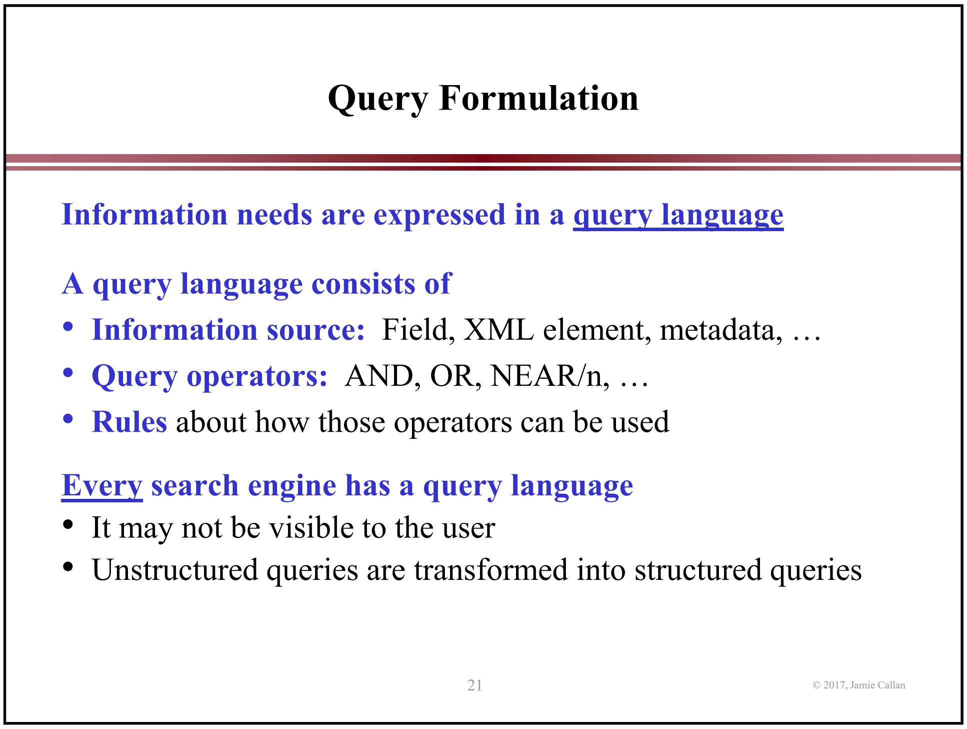
Production Request 56: Please produce any and all documents concerning soil water management as it pertains to commercial irrigation

Negotiated Query: (((Soil! OR sewage OR sewer! OR septic OR drain! OR dirt OR field! OR groundwater OR (ground w/3 water)) AND (manage! OR "control system")) AND irrigat!)

Query language details

- ! matches different stems (e.g., soil, soils, soiled, …)
- w/3 is NEAR/3
- “ ” is a phrase operator

(TREC 2007 Legal Track)



Query Formulation

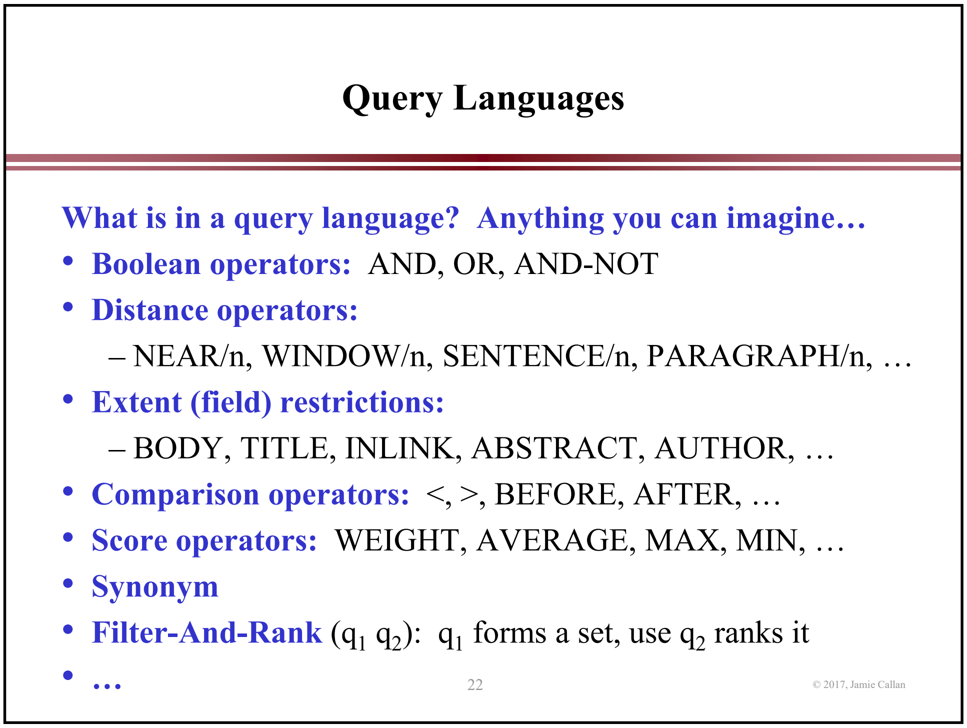
Information needs are expressed in a query language

A query language consists of

- **Information source:** Field, XML element, metadata, …
- **Query operators:** AND, OR, NEAR/n, …
- **Rules** about how those operators can be used

Every search engine has a query language

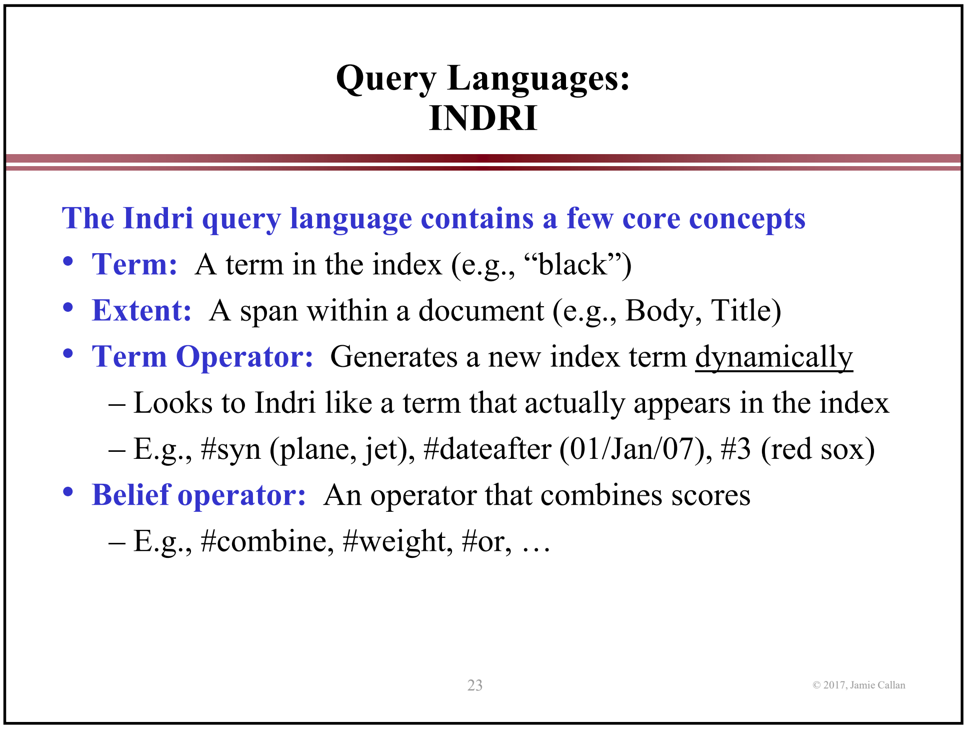
- It may not be visible to the user
- Unstructured queries are transformed into structured queries



Query Languages

What is in a query language? Anything you can imagine…

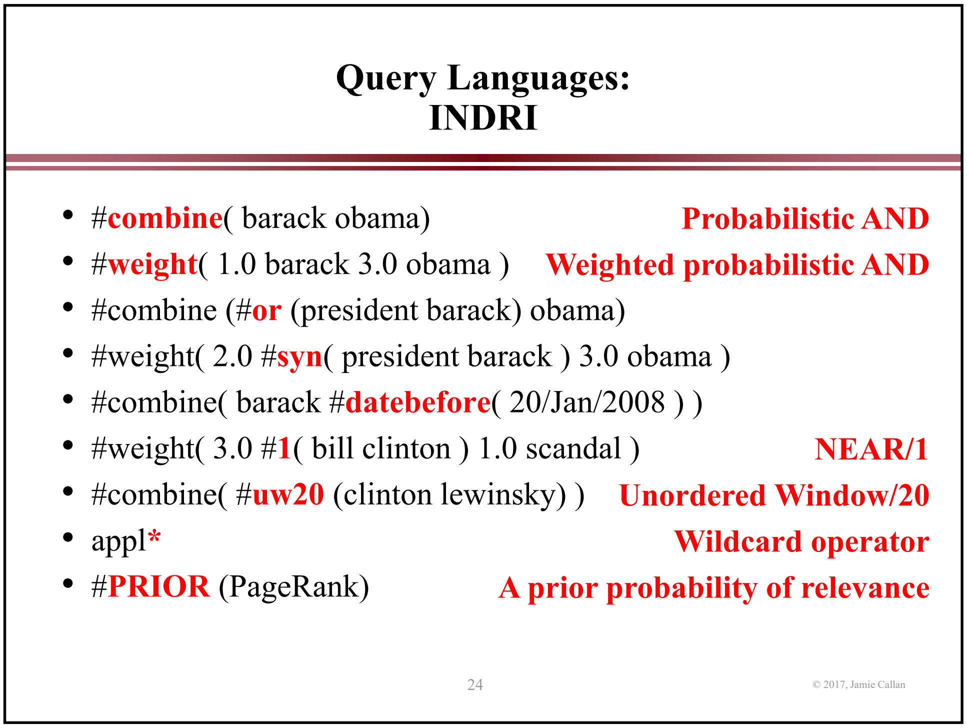
- **Boolean operators:** AND, OR, AND-NOT
- **Distance operators:**
  - – NEAR/n, WINDOW/n, SENTENCE/n, PARAGRAPH/n, …
- **Extent (field) restrictions:**
  - – BODY, TITLE, INLINK, ABSTRACT, AUTHOR, …
- **Comparison operators:** <, >, BEFORE, AFTER, …
- **Score operators:** WEIGHT, AVERAGE, MAX, MIN, …
- **Synonym**
- **Filter-And-Rank** (q_1 q_2): q_1 forms a set, use q_2 ranks it
- …



Query Languages: INDRI

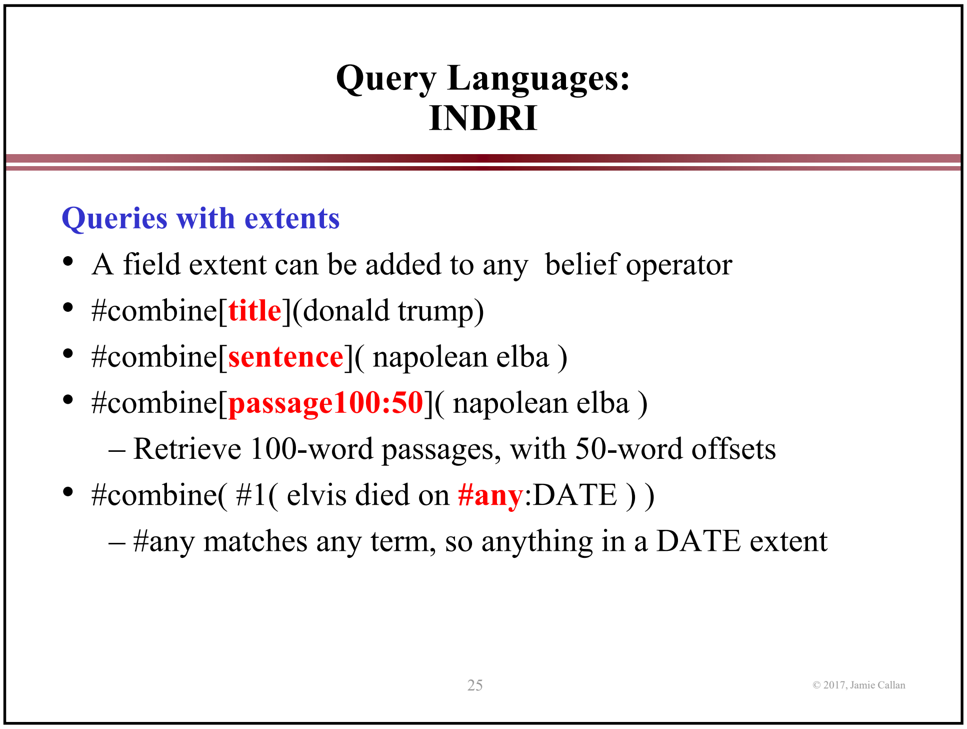
The Indri query language contains a few core concepts

- **Term:** A term in the index (e.g., “black”)
- **Extent:** A span within a document (e.g., Body, Title)
- **Term Operator:** Generates a new index term dynamically
  - Looks to Indri like a term that actually appears in the index
  - E.g., #syn (plane, jet), #dateafter (01/Jan/07), #3 (red sox)
- **Belief operator:** An operator that combines scores
  - E.g., #combine, #weight, #or, …



Query Languages: INDRI

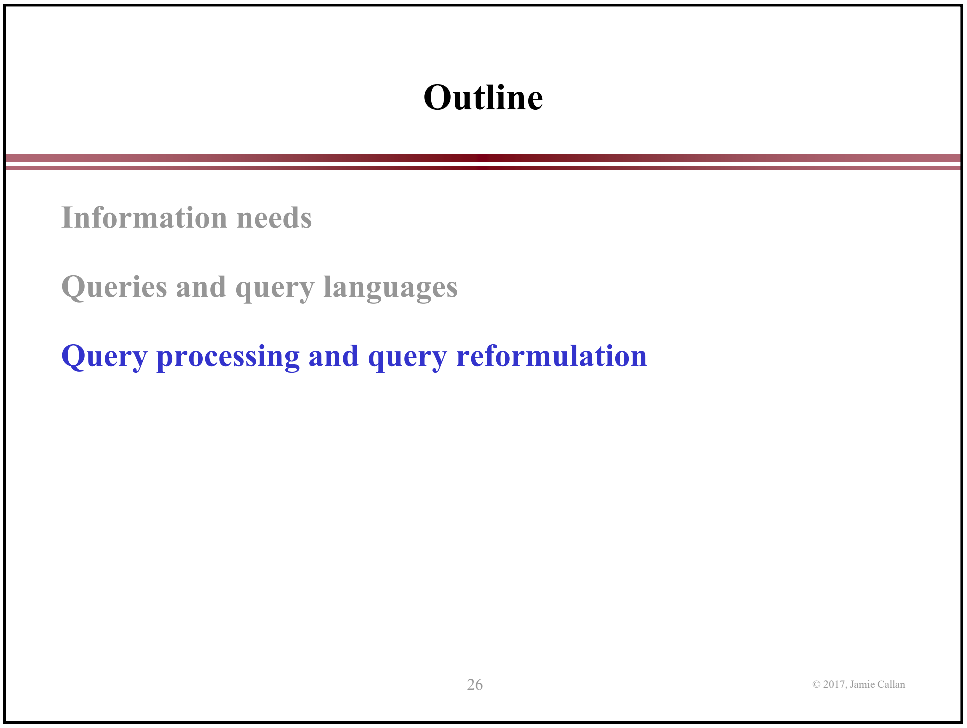
- #**combine**(barack obama) **Probabilistic AND**
- #**weight**(1.0 barack 3.0 obama) **Weighted probabilistic AND**
- #combine (#**or** (president barack) obama)
- #weight(2.0 #**syn**(president barack) 3.0 obama)
- #combine(barack #**datebefore**(20/Jan/2008))
- #weight(3.0 #**1**(bill clinton) 1.0 scandal) **NEAR/1**
- #combine(#**uw20** (clinton lewinsky)) **Unordered Window/20**
- appl***** **Wildcard operator**
- #**PRIOR** (PageRank) **A prior probability of relevance**



Query Languages: INDRI

Queries with extents

- A field extent can be added to any belief operator
- #combine[**title**](donald trump)
- #combine[**sentence**](napolean elba)
- #combine[**passage100:50**](napolean elba)
  - Retrieve 100-word passages, with 50-word offsets
- #combine(#1(elvis died on **#any**:DATE))
  - #any matches any term, so anything in a DATE extent

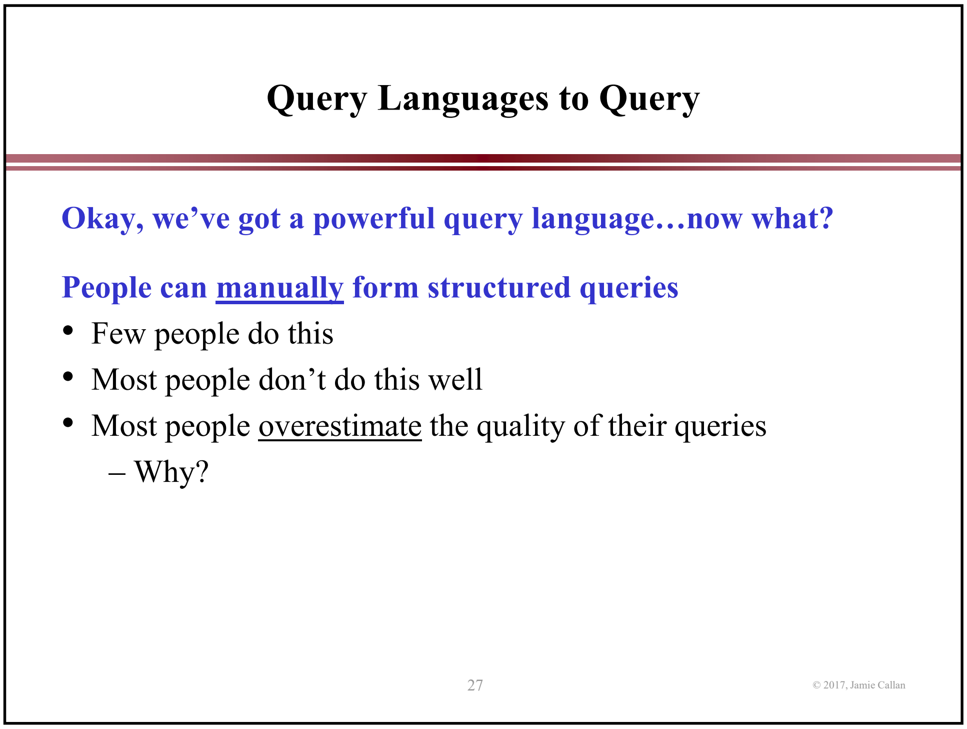


Outline

Information needs

Queries and query languages

Query processing and query reformulation

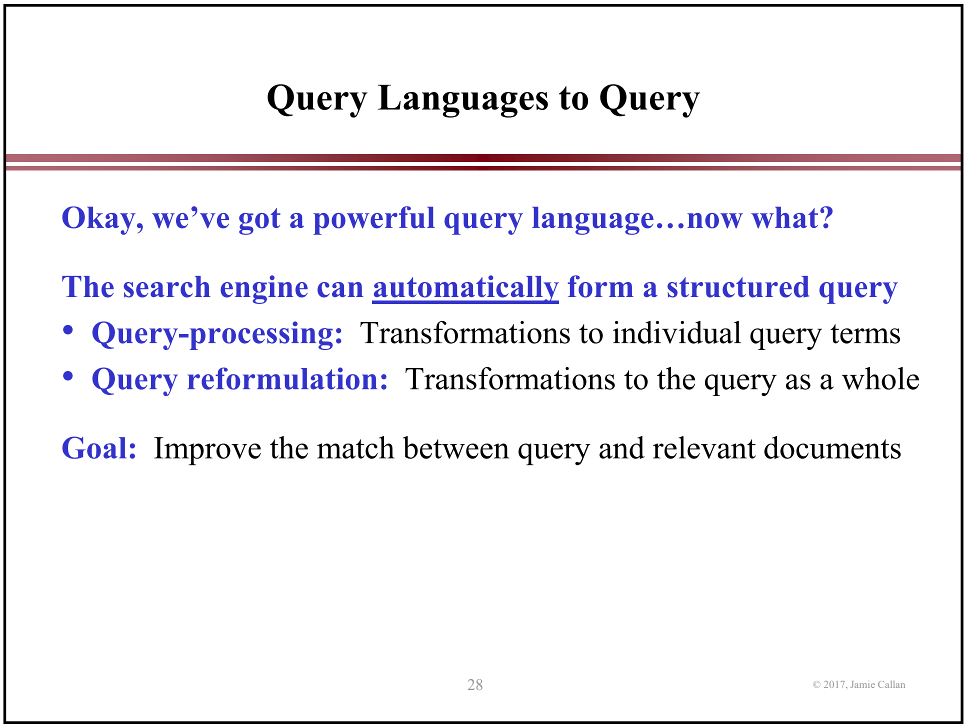


Query Languages to Query

Okay, we've got a powerful query language…now what?

People can manually form structured queries

- Few people do this
- Most people don't do this well
- Most people overestimate the quality of their queries
  - Why?



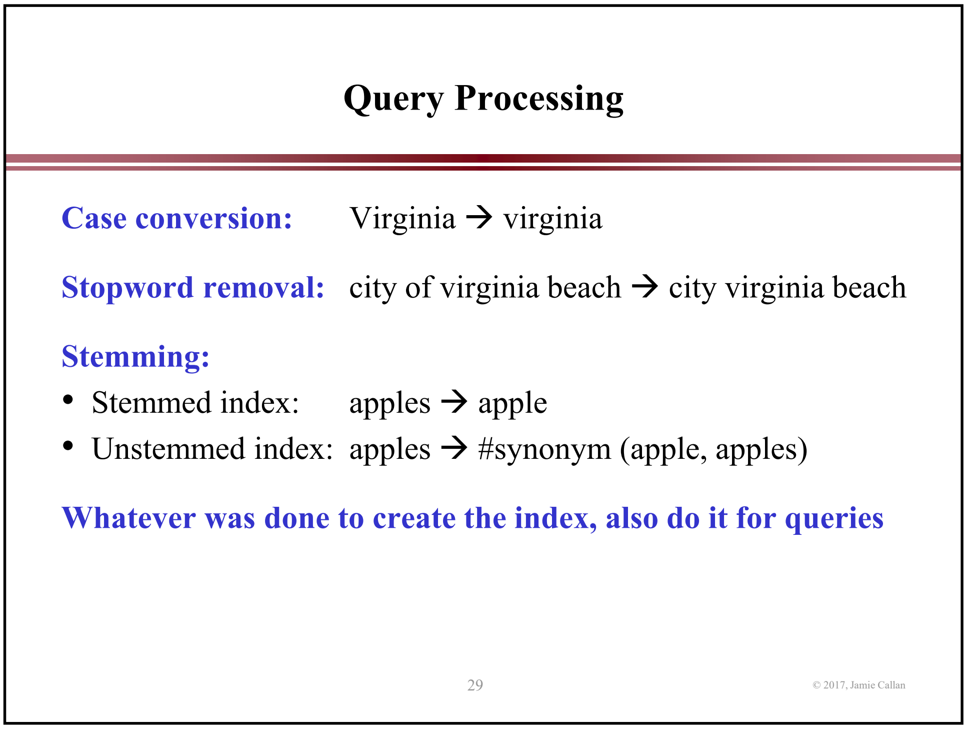
Query Languages to Query

Okay, we've got a powerful query language…now what?

The search engine can automatically form a structured query

- **Query-processing:** Transformations to individual query terms
- **Query reformulation:** Transformations to the query as a whole

Goal: Improve the match between query and relevant documents



Query Processing

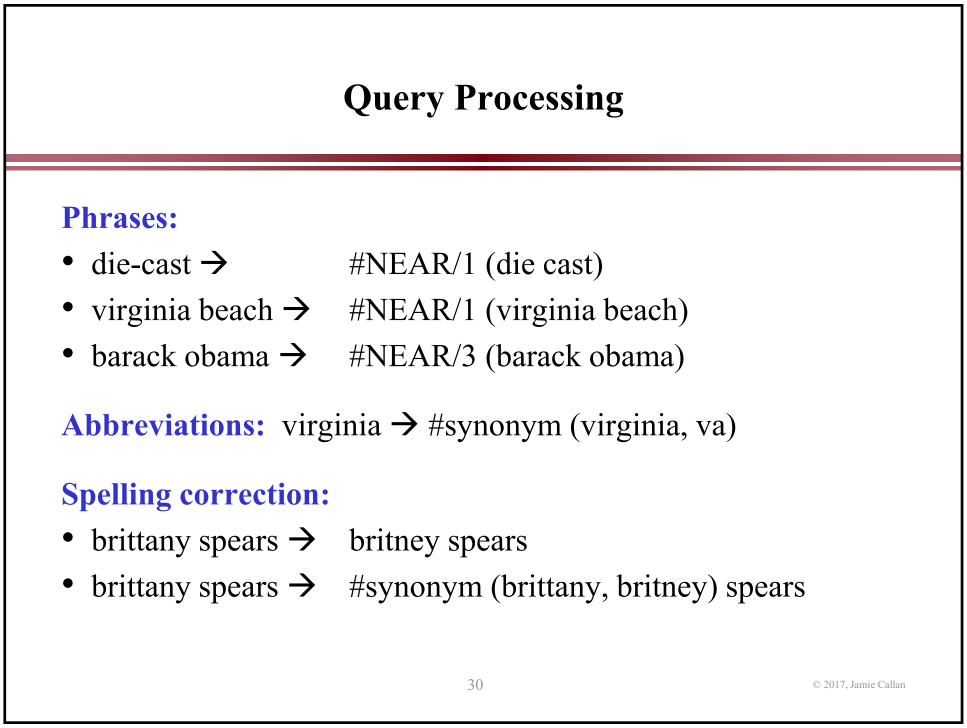
Case conversion: Virginia → virginia

Stopword removal: city of virginia beach → city virginia beach

Stemming:

- Stemmed index: apples → apple
- Unstemmed index: apples → #synonym (apple, apples)

Whatever was done to create the index, also do it for queries



Query Processing

Phrases:

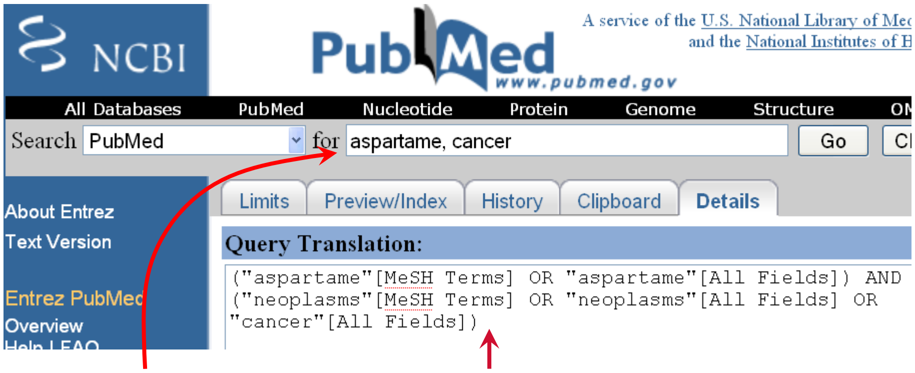
- die-cast → #NEAR/1 (die cast)
- virginia beach → #NEAR/1 (virginia beach)
- barack obama → #NEAR/3 (barack obama)

Abbreviations: virginia → #synonym (virginia, va)

Spelling correction:

- brittany spears → britney spears
- brittany spears → #synonym (brittany, britney) spears

Query Reformulation: Multiple Representations



Query Translation:

```
("aspartame"[MeSH Terms] OR "aspartame"[All Fields]) AND
("neoplasms"[MeSH Terms] OR "neoplasms"[All Fields] OR
"cancer"[All Fields])
```

The user typed an unstructured query

The system converted it to a structured query

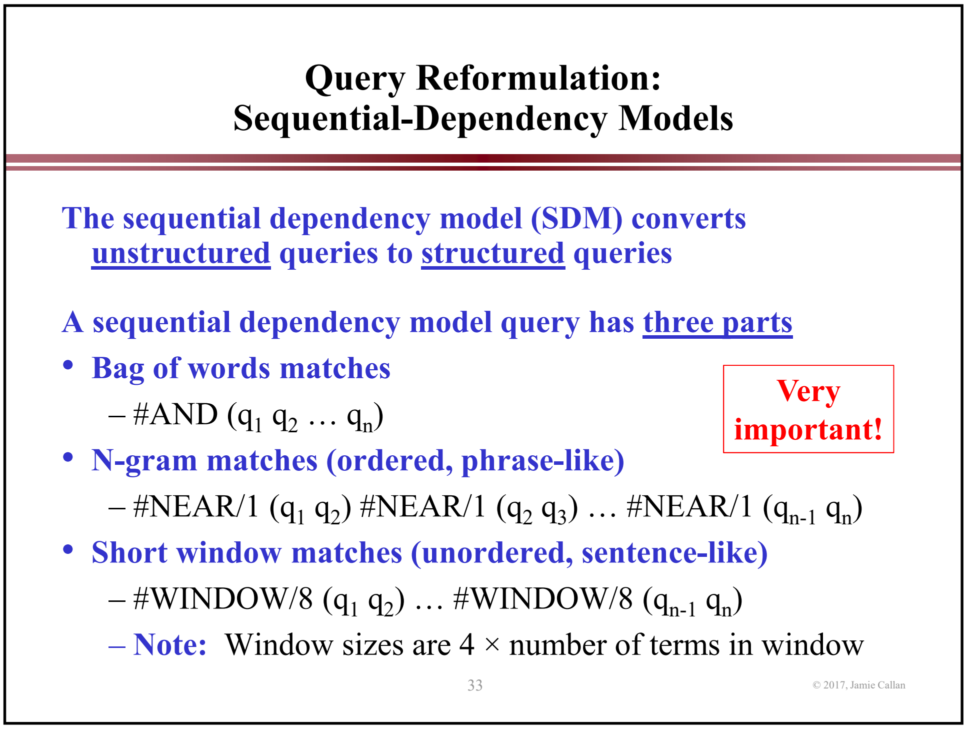
- Boolean and field operators
- Full text and controlled vocabulary terms

Query Reformulation: Multiple Representations

User query: The Time Traveler's Wife Search

A search engine might transform it into something like this

```
#and (
   #wsum(0.1 time.url       0.2 time.title       0.3 time.inlink
            0.4 time.body)
   #wsum(0.1 traveler.url  0.2 traveler.title  0.3 traveler.inlink
            0.4 traveler.body)
   #wsum(0.1 wife.url       0.2 wife.title       0.3 wife.inlink
            0.4 wife.body))
```



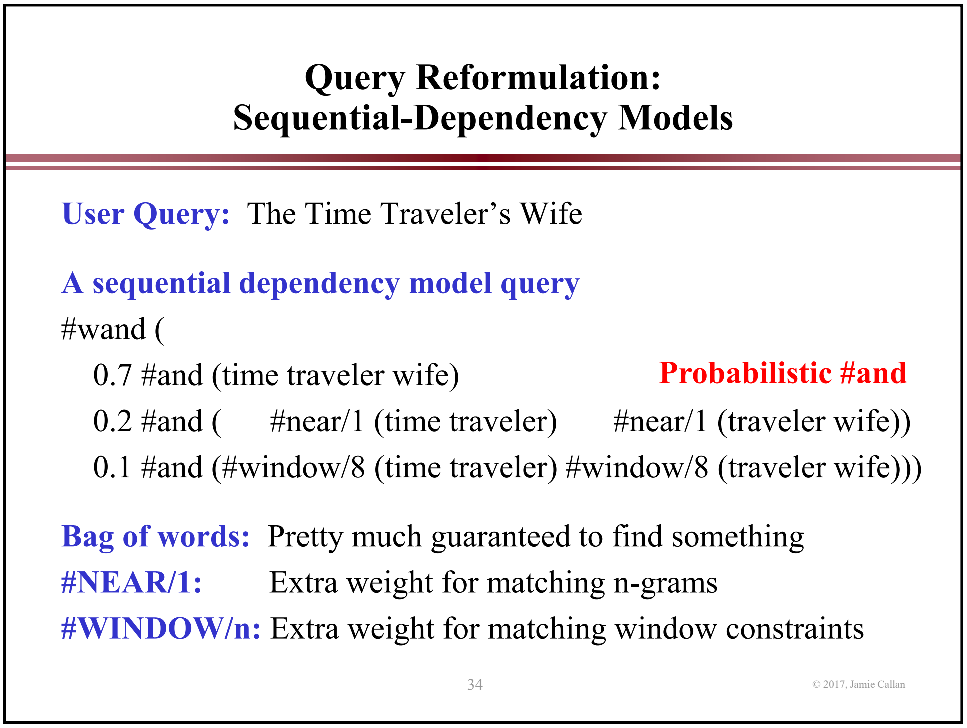
Query Reformulation: Sequential-Dependency Models

The sequential dependency model (SDM) converts <u>unstructured</u> queries to <u>structured</u> queries

A sequential dependency model query has <u>three parts</u>

- **Bag of words matches**
  - #AND (q_1 q_2 … q_n)
- **N-gram matches (ordered, phrase-like)**
  - #NEAR/1 (q_1 q_2) #NEAR/1 (q_2 q_3) … #NEAR/1 (q_{n-1} q_n)
- **Short window matches (unordered, sentence-like)**
  - #WINDOW/8 (q_1 q_2) … #WINDOW/8 (q_{n-1} q_n)
  - **Note:** Window sizes are 4 × number of terms in window

Very important!



Query Reformulation: Sequential-Dependency Models

User Query: The Time Traveler's Wife

A sequential dependency model query

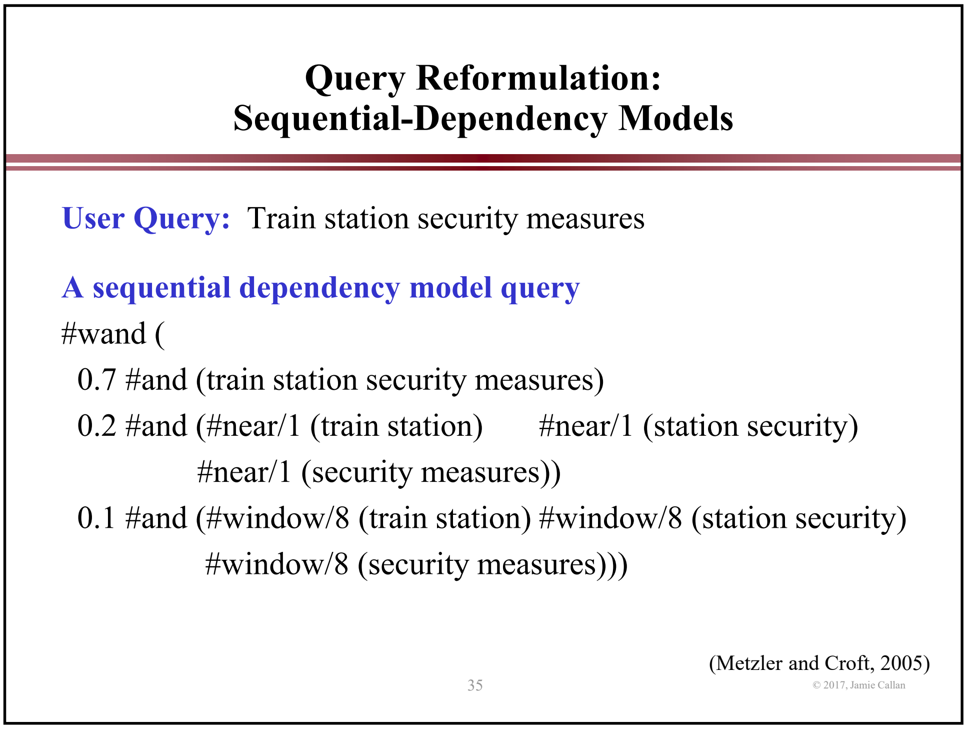
```
#wand (
   0.7 #and (time traveler wife)
   0.2 #and (     #near/1 (time traveler)     #near/1 (traveler wife))
   0.1 #and (#window/8 (time traveler) #window/8 (traveler wife)))
```

Probabilistic #and

Bag of words: Pretty much guaranteed to find something

#NEAR/1: Extra weight for matching n-grams

#WINDOW/n: Extra weight for matching window constraints



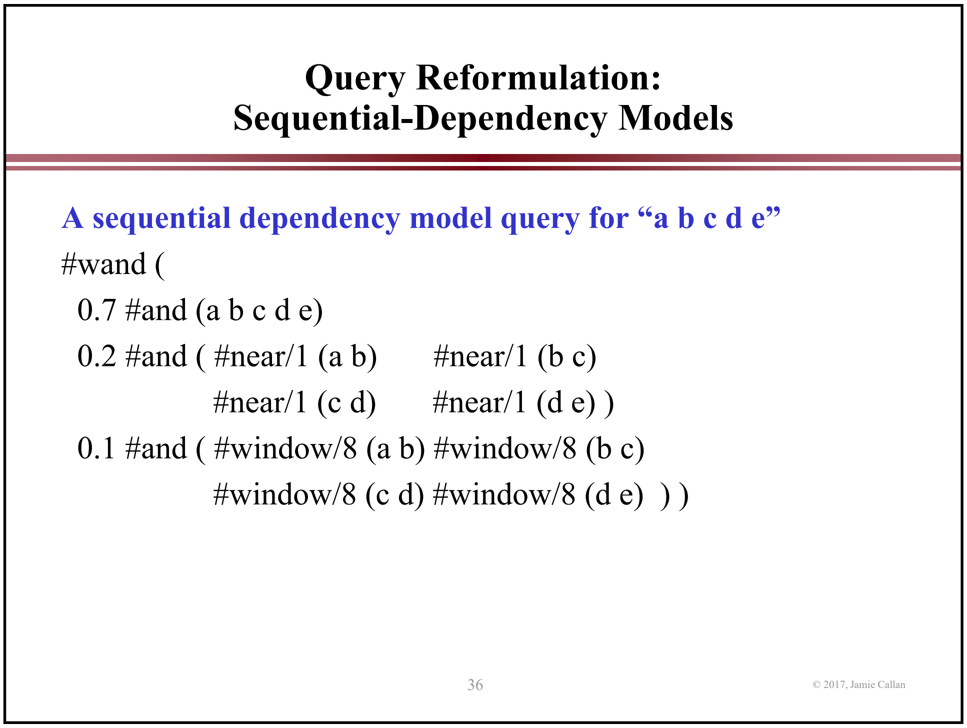
Query Reformulation: Sequential-Dependency Models

User Query: Train station security measures

A sequential dependency model query

```
#wand (
  0.7 #and (train station security measures)
  0.2 #and (#near/1 (train station)      #near/1 (station security)
               #near/1 (security measures))
  0.1 #and (#window/8 (train station) #window/8 (station security)
                #window/8 (security measures)))
```

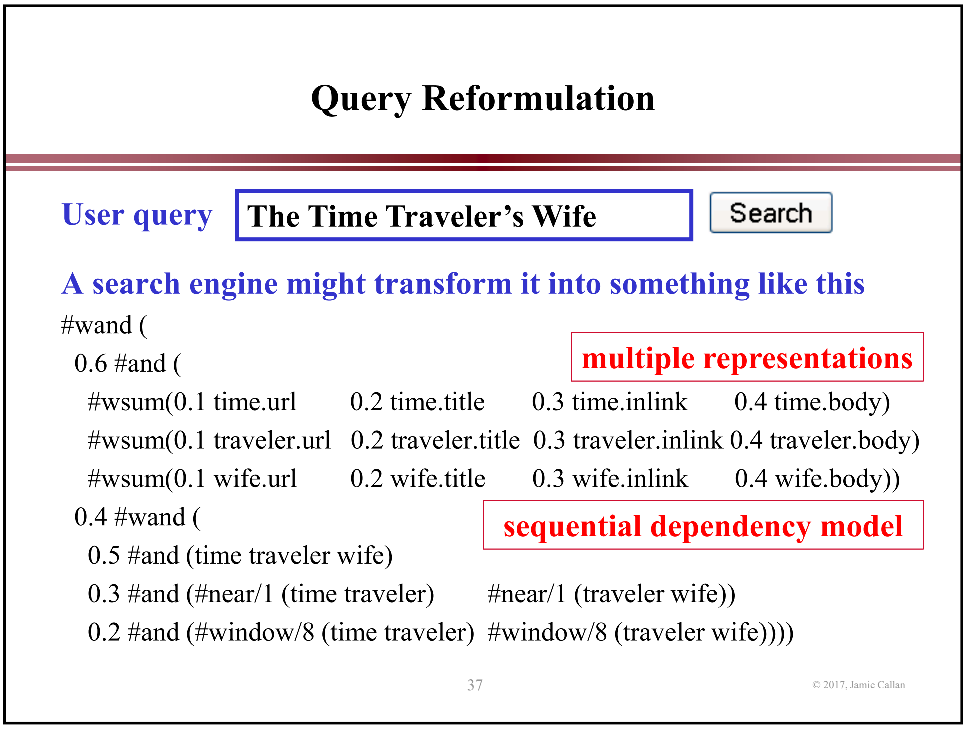
(Metzler and Croft, 2005)



Query Reformulation: Sequential-Dependency Models

A sequential dependency model query for "a b c d e"

```
#wand (
  0.7 #and (a b c d e)
  0.2 #and ( #near/1 (a b)      #near/1 (b c)
                #near/1 (c d)      #near/1 (d e) )
  0.1 #and ( #window/8 (a b) #window/8 (b c)
                #window/8 (c d) #window/8 (d e)  ) )
```

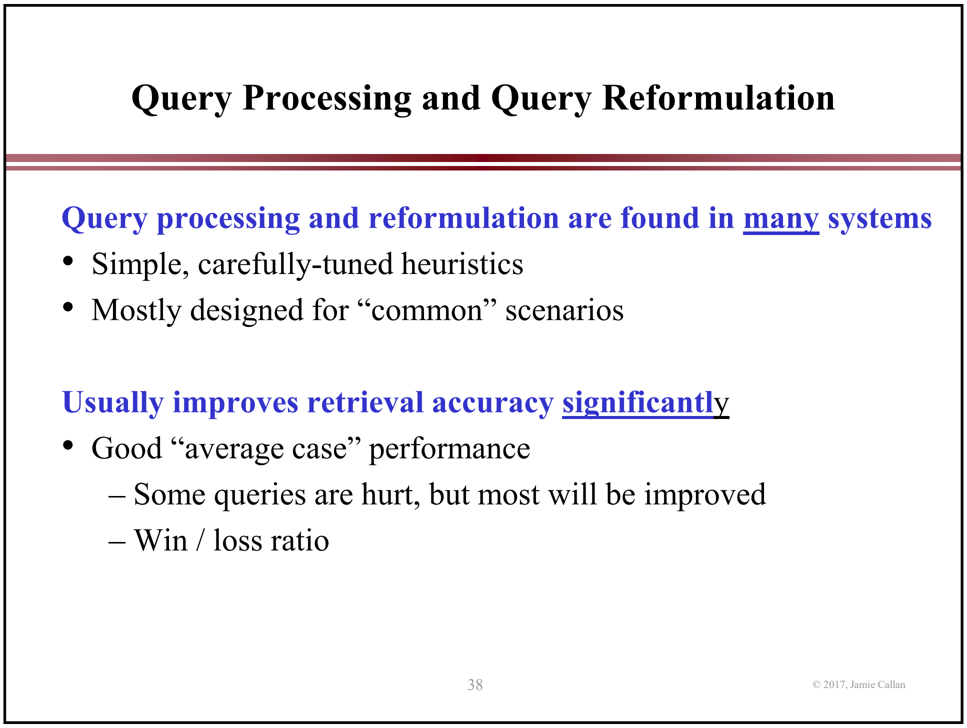


Query Reformulation

User query **The Time Traveler's Wife** Search

A search engine might transform it into something like this

```
#wand (
  0.6 #and (                                        multiple representations
    #wsum(0.1 time.url       0.2 time.title      0.3 time.inlink      0.4 time.body)
    #wsum(0.1 traveler.url   0.2 traveler.title  0.3 traveler.inlink 0.4 traveler.body)
    #wsum(0.1 wife.url       0.2 wife.title      0.3 wife.inlink      0.4 wife.body))
  0.4 #wand (                                 sequential dependency model
    0.5 #and (time traveler wife)
    0.3 #and (#near/1 (time traveler)        #near/1 (traveler wife))
    0.2 #and (#window/8 (time traveler)  #window/8 (traveler wife))))
```



Query Processing and Query Reformulation

Query processing and reformulation are found in many systems

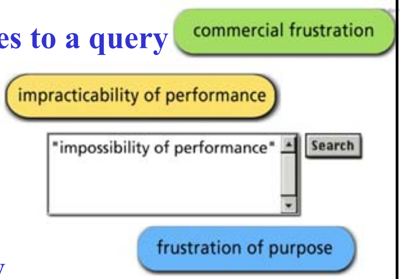
- Simple, carefully-tuned heuristics
- Mostly designed for "common" scenarios

Usually improves retrieval accuracy significantly

- Good "average case" performance
  - – Some queries are hurt, but most will be improved
  - – Win / loss ratio

Query Reformulation: Query Expansion

Query expansion adds words / phrases to a query


commercial frustration
impracticability of performance
"impossibility of performance"
Search
frustration of purpose

WestLaw

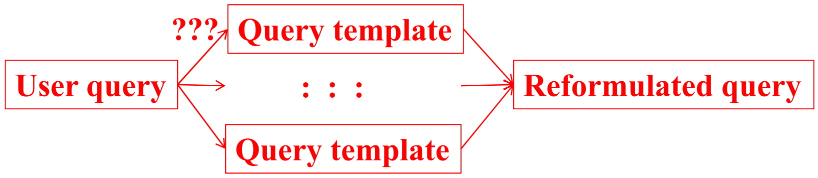
Sources of words / phrases:

- A full-text synonym dictionary
  - Not a general thesaurus
- A controlled vocabulary dictionary
- Text mine in retrieved documents
  - Average case performance is good, so researchers like it
  - High variance, so not used much in operational systems
- Text mine in query logs

Query Classification and Reformulation

There are different types of information needs…

- Use classification to identify the query intent
  - And typically associates it with special-purpose processing



???

Query template

User query

: : :

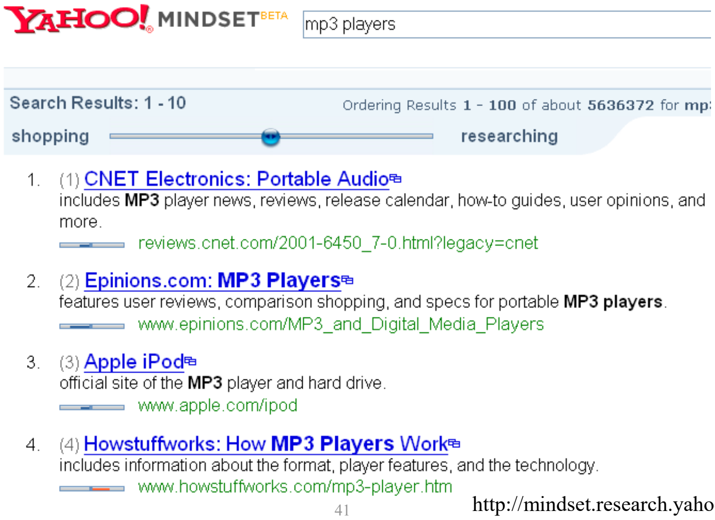
Reformulated query

Query template

- This is done by Web search engines
  - But not a lot is known about how they do it
- We will cover this in a later lecture

(E.g., U.S. Patent 20060190439)

A Typical Document Ranking



YAHOO! MINDSET BETA mp3 players

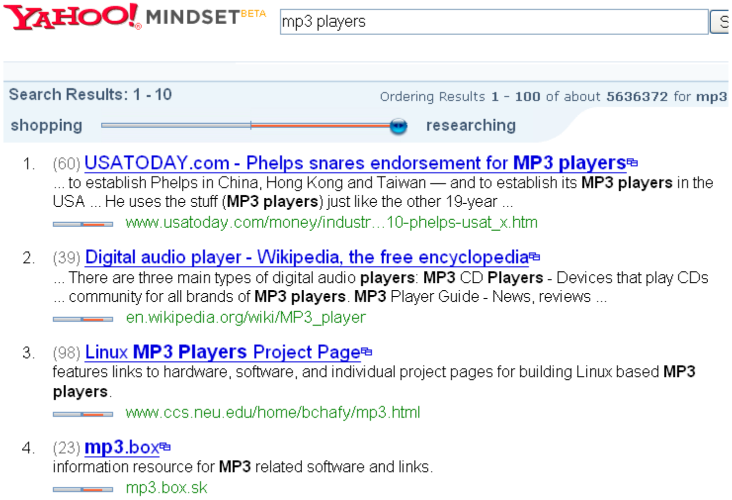
Search Results: 1 - 10 Ordering Results **1 - 100** of about **5636372** for **mp**

shopping researching

1. (1) CNET Electronics: Portable Audio
includes **MP3** player news, reviews, release calendar, how-to guides, user opinions, and more.
reviews.cnet.com/2001-6450_7-0.html?legacy=cnet
2. (2) Epinions.com: **MP3 Players**
features user reviews, comparison shopping, and specs for portable **MP3 players**.
www.epinions.com/MP3_and_Digital_Media_Players
3. (3) Apple iPod
official site of the **MP3** player and hard drive.
www.apple.com/ipod
4. (4) Howstuffworks: How **MP3 Players** Work
includes information about the format, player features, and the technology.
www.howstuffworks.com/mp3-player.htm

http://mindset.research.yahoo.com/

A Document Ranking Optimized For Informational Queries



YAHOO! MINDSET BETA mp3 players S

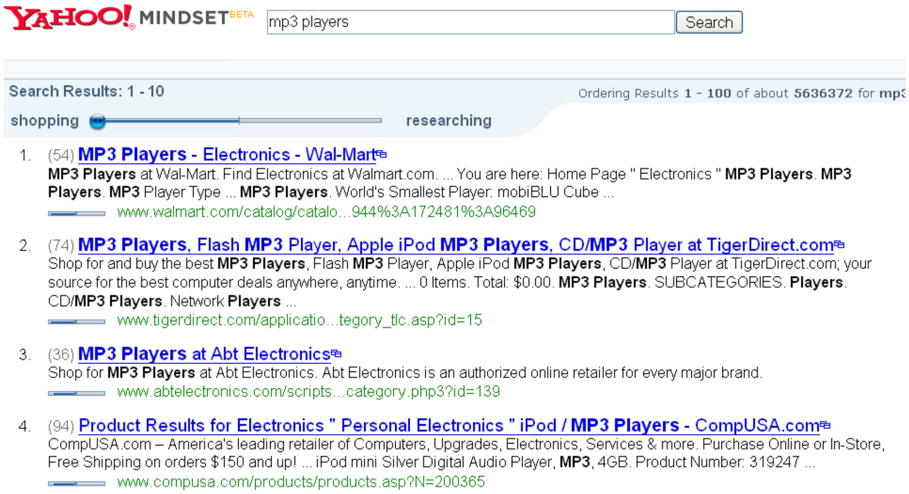
Search Results: 1 - 10 Ordering Results **1 - 100** of about **5636372** for **mp3**

shopping researching

1. (60) USATODAY.com - Phelps snares endorsement for **MP3 players**
... to establish Phelps in China, Hong Kong and Taiwan — and to establish its **MP3 players** in the USA ... He uses the stuff (**MP3 players**) just like the other 19-year ...
www.usatoday.com/money/industr…10-phelps-usat_x.htm
2. (39) Digital audio player - Wikipedia, the free encyclopedia
... There are three main types of digital audio **players**: **MP3** CD **Players** - Devices that play CDs ... community for all brands of **MP3 players**. **MP3** Player Guide - News, reviews ...
en.wikipedia.org/wiki/MP3_player
3. (98) Linux **MP3 Players** Project Page
features links to hardware, software, and individual project pages for building Linux based **MP3 players**.
www.ccs.neu.edu/home/bchafy/mp3.html
4. (23) **mp3**.box
information resource for **MP3** related software and links.
mp3.box.sk

http://mindset.research.yahoo.com/

A Document Ranking Optimized For Transaction Queries



Search Results: 1 - 10

Ordering Results **1 - 100** of about **5636372** for **mp3**

1. (54) **MP3 Players** - Electronics - Wal-Mart
   MP3 Players at Wal-Mart. Find Electronics at Walmart.com. ... You are here: Home Page " Electronics " **MP3 Players**. **MP3 Players**. **MP3** Player Type ... **MP3 Players**. World's Smallest Player: mobiBLU Cube ...
   www.walmart.com/catalog/catalo...944%3A172481%3A96469
2. (74) **MP3 Players**, Flash **MP3** Player, Apple iPod **MP3 Players**, CD/**MP3** Player at TigerDirect.com
   Shop for and buy the best **MP3 Players**, Flash **MP3** Player, Apple iPod **MP3 Players**, CD/**MP3** Player at TigerDirect.com; your source for the best computer deals anywhere, anytime. ... 0 Items. Total: $0.00. **MP3 Players**. SUBCATEGORIES. **Players**. CD/**MP3 Players**. Network **Players** ...
   www.tigerdirect.com/applicatio...tegory_tlc.asp?id=15
3. (36) **MP3 Players** at Abt Electronics
   Shop for **MP3 Players** at Abt Electronics. Abt Electronics is an authorized online retailer for every major brand.
   www.abtelectronics.com/scripts...category.php3?id=139
4. (94) Product Results for Electronics " Personal Electronics " iPod / **MP3 Players** - CompUSA.com
   CompUSA.com – America's leading retailer of Computers, Upgrades, Electronics, Services & more. Purchase Online or In-Store, Free Shipping on orders $150 and up! ... iPod mini Silver Digital Audio Player, **MP3**, 4GB. Product Number: 319247 ...
   www.compusa.com/products/products.asp?N=200365

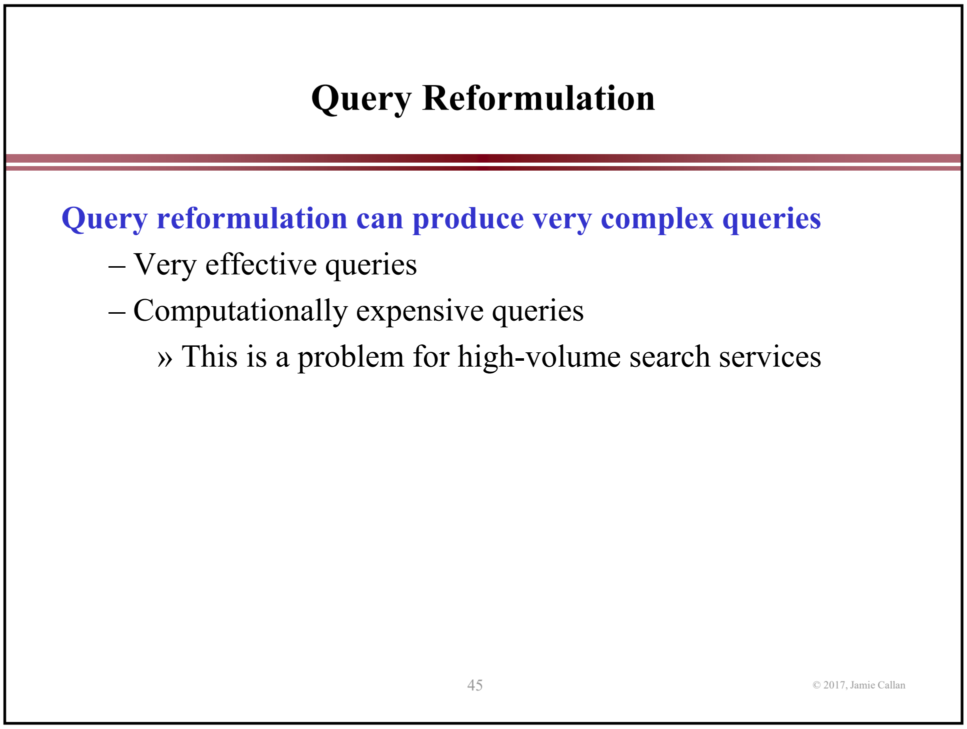
http://mindset.research.yahoo.com/

Query Reformulation on the Web

Jon Pederson (Bing) says…

- **Query understanding is critical to web search**
  - Affects most queries
  - Can radically improve results
- **Trade-off between relevance and efficiency**
  - Rewrites can be costly
  - Win/loss ratio is the key metric
- **Especially important for tail queries**
  - No meta-data to guide matching and ranking

(Pederson, 2010)

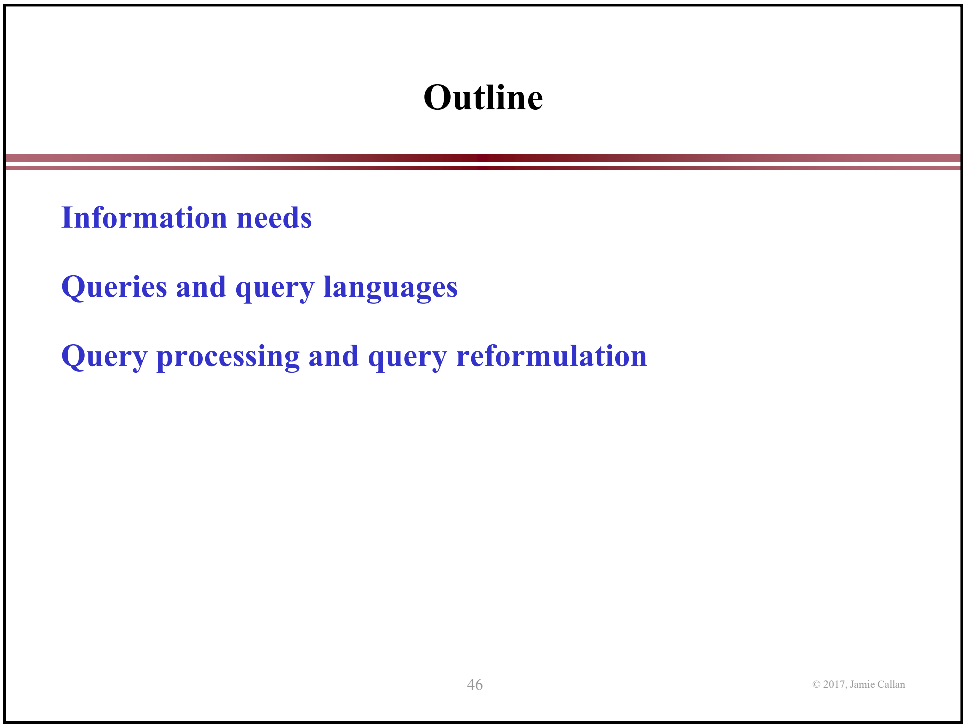


Query Reformulation

Query reformulation can produce very complex queries

- – Very effective queries
- – Computationally expensive queries
  - » This is a problem for high-volume search services

45



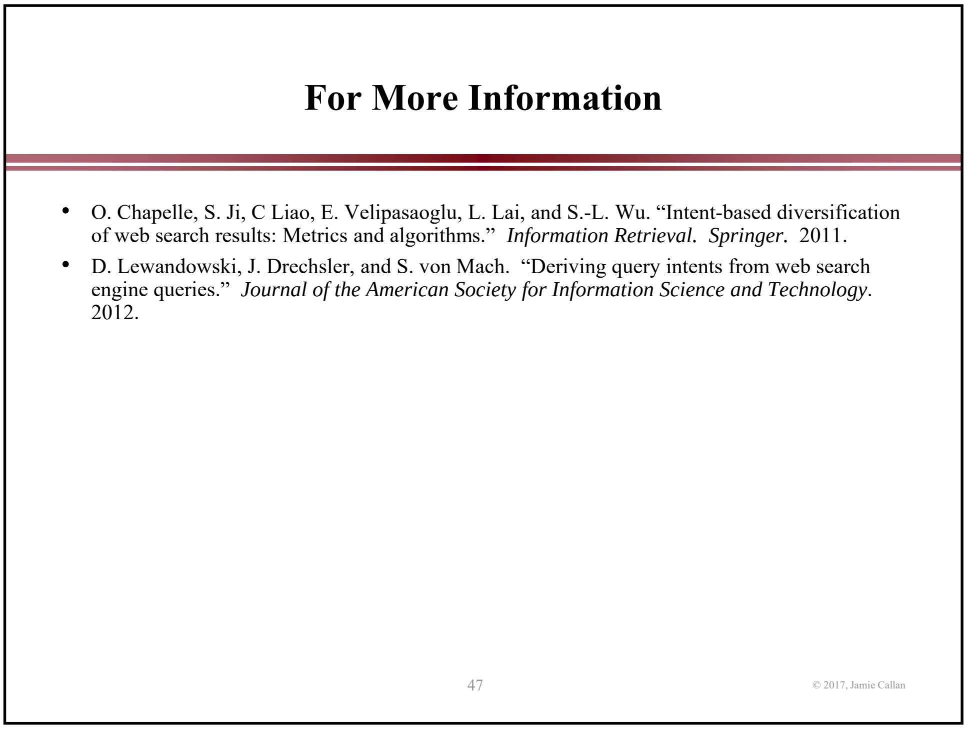
Outline

Information needs

Queries and query languages

Query processing and query reformulation

46



For More Information

- O. Chapelle, S. Ji, C Liao, E. Velipasaoglu, L. Lai, and S.-L. Wu. “Intent-based diversification of web search results: Metrics and algorithms.” *Information Retrieval. Springer.* 2011.
- D. Lewandowski, J. Drechsler, and S. von Mach. “Deriving query intents from web search engine queries.” *Journal of the American Society for Information Science and Technology*. 2012.