Introduction to Information Retrieval

School of Information Studies
Syracuse University

What Is Information Retrieval (IR)

Gerard Salton, 1968:
Information retrieval is a field concerned with the structure, analysis, organization, storage, searching and retrieval of information.

Manning, Raghavan, and Schutze, 2008:
Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

- "Document" is the generic term for an information holder (book, chapter, article, webpage, etc.)

Web search is the branch of IR where the collection of documents includes those that are located on the Web.

What Is Tough About IR?

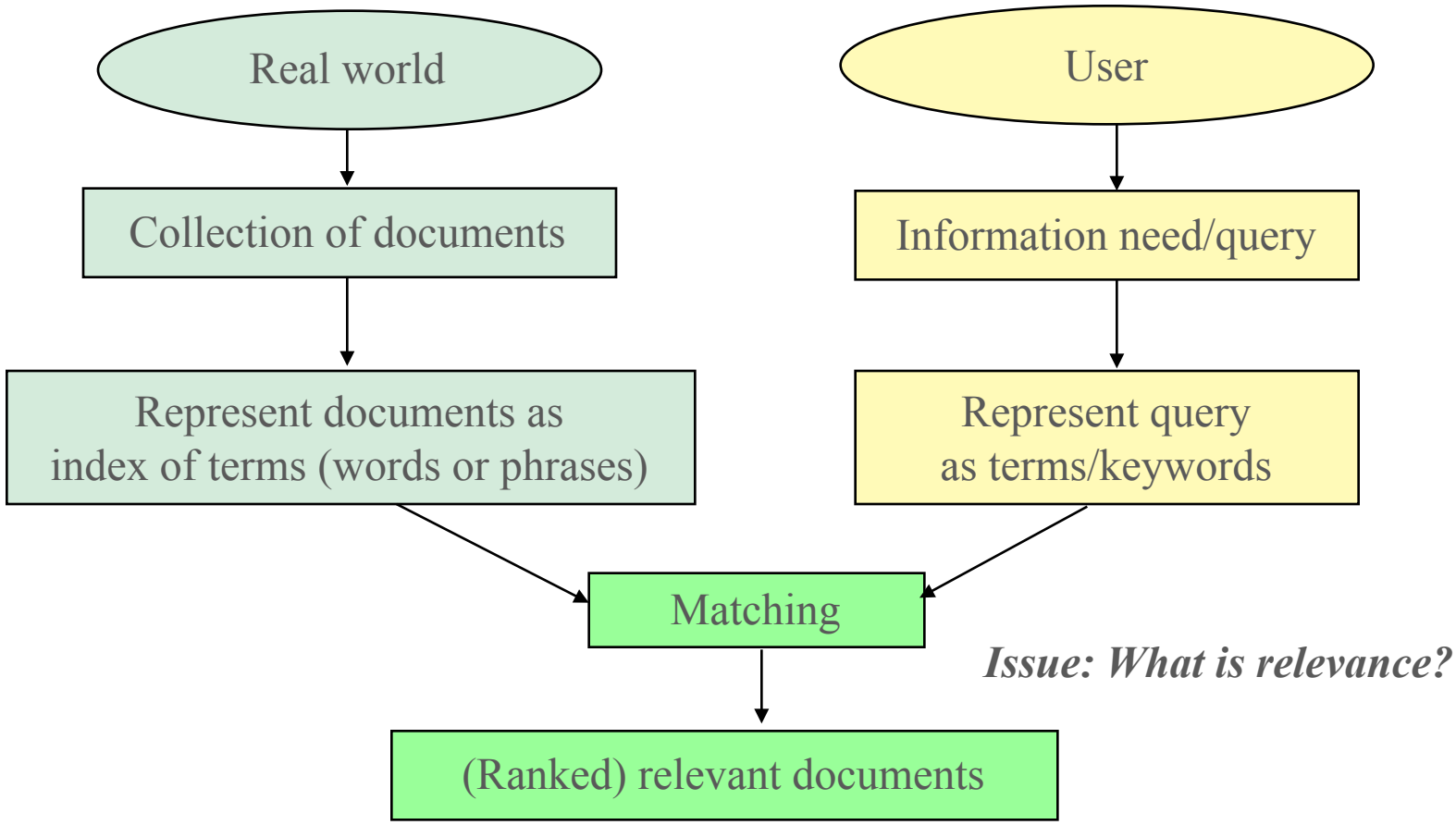
One issue is **how to represent documents** so that others might retrieve them.

- Need to match the text of the document with the query
- In full, free-text systems, this is an issue because documents and queries are expressed in language
  - And language is synonymous and polysemous
  - Methods for solving the language issue are difficult
- Sometimes called the **vocabulary gap** or mismatch

Given the retrieval of some documents, how do we decide which ones are **most relevant** to the user's query?

- Most often implemented as a **ranking** of the resulting documents

Typical Information Retrieval System



Text Retrieval Conference (TREC)

Co-sponsored by the National Institute of Standards and Technology (NIST) and the Defense Advanced Research Projects Agency (DARPA) and begun in 1992

Purpose is to **support research within the information-retrieval community** by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies

- Provides document collections, queries, and human judges
- Main IR track was called the "ad-hoc retrieval track"

Has grown in the number of participating systems and the number of tracks each year

- Tracks have included cross-language retrieval, filtering, question answering, interactive, Web, novelty, video, blog search…

IR System Research

Traditional IR system research assumes that a user is interested in finding out information on a particular topic

- Idea of the user is a student investigating documents for a paper, or a researcher interested in a topic

TREC collections and research experiments

- Build IR systems with different retrieval models
- Test against a standard collection of newswire documents
- Human evaluators judge relevant documents
- Report system evaluations in terms of precision and recall
- Example type of query:
  I am interested in all documents that discuss oil reserves and current attempts to find new reserves, particularly those that discuss the international financial aspects of the oil-production process.

Information Needs

Other branches of research focus on the user and whether the user's underlying information seeking is satisfied

Early theories by Belkin, Oddy, and others

- Functions of the retrieval system to model the user's information need in an interactive retrieval session
  - Characterize user
  - Get initial information need
  - Develop need context
  - Formulate information need
  - Conduct search for documents
  - Evaluate results
  - Feedback from user

Information Retrieval (IR) Systems

School of Information Studies
Syracuse University

IR Systems: Constructing the Index

Process documents and identify terms to be indexed.

- Terms are often just the words
  - Usually stemming is applied and stop words removed
- Sometimes basic noun phrases are also added, particularly proper names

Compute weights of terms, depending on model definition.

Build an index, a giant dictionary mapping terms to documents.

- For each term
  - Keep a list of documents in which it occurs
  - Weights

IR Systems: Models

Vector space models

- Widely used weights known as TF-IDF (term frequency * inverted document frequency), TF-IDF = TF * IDF
  - **TF: frequency of the term** in the document (normalized by document length)—freqtd/ lengthd
    - Intuition: more frequently occurring terms are more important
  - **IDF: invert the document frequency**, the number of documents in the collection that the term occurs in—log(*N*/*nt*), where *N* is number of docs, *nt* is number of docs with term *t*
    - Intuition: terms occurring in all documents are less important to distinguish which ones are relevant to the query

Other models

- Probabilistic models
- Language models
- Boolean models

IR Systems: Queries and matching

Natural language queries are converted to terms, usually called keywords.

- In a Web search, typical queries are keywords already.

Query terms are used to retrieve documents from the index.

The model defines how to match query terms to documents, using the weights, and usually resulting in a score for each document.

Documents are returned in order of relevance score.

IR Systems: Evaluation

Human judgments as to whether returned documents are relevant to the query

Precision and recall can be used to evaluate a set of returned documents

Human judgments -> System:	**Relevant**	**Non-Relevant**
Retrieved	*a* (true positives)	*b* (false positives)
Non-Retrieved	*c* (false negatives)	*d* (true negatives)

Precision = $a/a + b$

Recall = $a/a + c$

IR Systems: Other Evaluation Measures

The **F-measure** is a combination of recall and precision, averaged using the harmonic mean

- Let P be precision and R be recall

$$F = (\beta 2 + 1)\, PR/(\beta 2)\, P + R$$

- Typically, the measure is used for $\beta = 1$, giving equal weight to precision and recall

$$F_{\beta = 1} = 2\, PR/P + R$$

Ranked retrieval evaluation

- Given the top *k*-ranked documents, compute precision and recall at every position
- Mean average precision
  - Average the precisions over all positions *k* in the ranking

IR Systems: Improving Retrieval

Query expansion, adding semantically similar words or context words

- For example, use WordNet to add synonyms to query terms
  - What sense to use? The first?
- Results are mixed
  - Synonyms added for incorrect sense will throw results off badly

IR Systems: Improving Retrieval

Relevance feedback

- The one technique consistently shown to improve retrieval
- Human relevance feedback: after human has selected a few really relevant documents, add terms from those documents to the query
- Pseudo-relevance feedback
  - Perform one retrieval and assume that the top *n* documents are relevant.
  - Use those documents to add terms to the query.

Web Search

School of Information Studies
Syracuse University

Web Search

With the advent of the Web, basic IR was applied to this scenario of linked documents worldwide

- Company like Google keeps a giant index of documents for search

Why/how would IR be different on the Web?

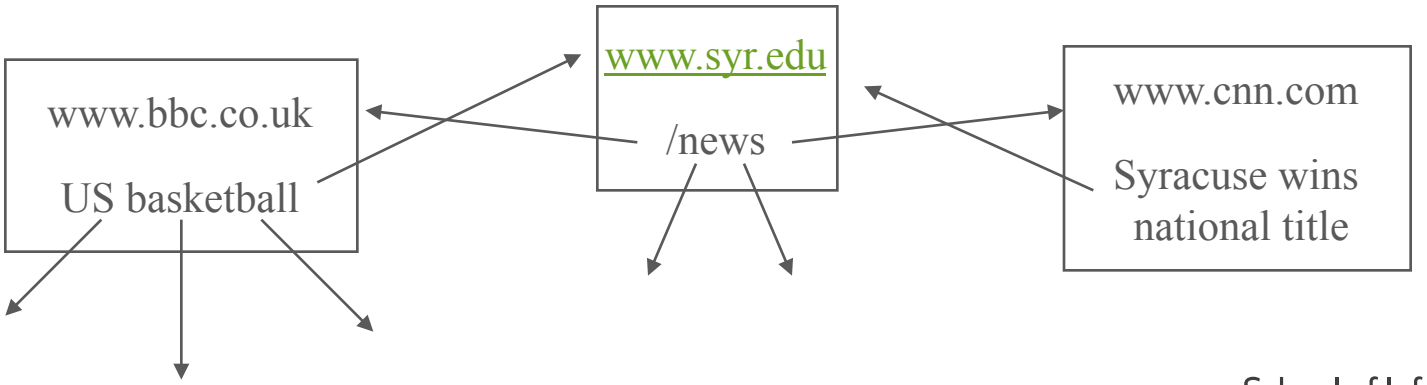
- The Web, compared with a database of documents
  - Is far larger
  - Is more dynamic—websites update all the time, links may not be permanent
  - Collection frequencies needed for inverse document frequency (IDF) are so impermanent
- Quality control of documents on the Web is not present
- No such thing as a complete inverted file for the entire Web—many hidden pages (deep Web)
- Importance of ranking results and the impact of pay for ranking

The Web Has Structure: Web Graph

View the collection of static webpages as a graph with "hyperlinks" between them

Hyperlink in HTML, given by the anchor tag, will give the URL of another webpage

- In-degree is the number of links coming to a page from other pages
- Out-degree is the number of links on the page



Building Search Engines: Web Crawling

In order to build an index of documents for Web search, the Web crawler, or spider, has to locate documents

Required features

- Robustness: it must not get stuck in dead ends or loops
- Politeness: it must not overwhelm any Web server with too much speed or too many requests
  - Web servers set politeness policies

Desired features

- Quality: should try to give "useful" pages priority
- Freshness: should obtain updated pages so that the Web index has a fairly current version of the webpage
- Performance and efficiency, scalability, operate in a distributed fashion

Building Search Engines: Web Document Processing

Find content and process into tokens for traditional use in IR indexing

- Content may be text in-between tags
- Image tags may have text attributes to describe the image
- May discard JavaScript and other computational elements
- May even try to discard "noisy" text in the form of website navigation, standard copyright notices, etc.
  - One technique is to observe that real-content text has fewer tags per token than non-content text

Keywords may be added to the document that don't appear directly in the content

- Metadata tags may have keywords
- Special weights may be added for tokens appearing in header tags
- Anchor text from other pages (see next slide)

Building Search Engines: Anchor Text

Sometimes the text content of a webpage does not contain generally descriptive words for that page.

- Home page for IBM did not contain the word "computer"
- Home page for Yahoo! did not contain the word "portal"

Generally descriptive words may be found in anchor text of links, or even near it, that occur in other pages.

 Big Blue

example of a large computing firm
 Big Blue

An example of a large computing firm is:
 here

- Typically, disregard anchor text words such as "click" and "here"
- Otherwise, we can add anchor text as keywords

Building Search Engines: Link Analysis

Link analysis can be viewed as a development of citation analysis for the Web.

- Bibliographic citation analysis used book and article references
- Bibliometric analysis of bibliographic citation links
  - Web examples: Web of Science from ISI/Citeseer

The intuition behind link analysis is that a hyperlink from page A to page B represents an endorsement of page B, by the creator of page A.

- Not true for some links, such as links to administrative notices on corporate websites—"internal" links are typically discounted

Building Search Engines: Link Analysis

Two major algorithms, **PageRank and HITS**, that give scoring weights for webpages

- PageRank is from Sergei and Brin, founders of Google
- Such weights are combined with other weights from content tokens and many other ranking criteria

Additional Criteria for Ranking

Popularity: What are the current topics of the day?

- Collected from blogs and previous queries

Click-through results: statistics about which pages users click on after getting ranked results can inform ranking algorithms to improve later rankings

Context: keep track of the user's interests, location, situation

- What do other users like this one like?

Learning to rank: use machine learning on ranked relevance results to improve rankings

- Importance of getting relevant documents in the top-10 list
- Search-engine companies have large amounts of data, including relevance judgments in terms of what documents users click on after a query

More on Web Search

School of Information Studies
Syracuse University

Evaluating Ranked Retrieval Results

Evaluation measure: **discounted cumulative gain (DCG)**

- Measures relevance at each ranked position
- Penalizes highly relevant documents that are lower down in the ranks
- DCG normalizes over queries (of different lengths)

Experiments for search engines

- User judgments are good but are necessarily small in scope
- A/B testing
  - Deploy an experimental search engine to some users (group B) while other users get "normal" search engine (group A)
  - Use "click-through" judgments as to which results the users thought would be relevant
  - Evaluate which relevant results are highest ranked

Where Is the NLP in IR?

IR is thought of as a field in its own right, but NLP is used at the lower levels in building indexes

- Stemming, stopwords
- Named entities and other noun phrases

Web search engines incorporating natural language and NLP into queries

- Not just keywords anymore
- Query processing to find commonly used patterns and tailor the searches and
  search results
  - Conversion of units
  - Queries with "How do I…" or other patterns
  - Other query-processing techniques similar to those in question/answering

Search Engine Company Data Centers

Google designs data centers specifically for Web search.

- http://www.google.com/about/datacenters/
- Uses a lot of low-cost computers networked together
  - Design network and data algorithms for fast performance
- Acknowledges 15 data centers around the world (probably two dozen more)
  - Each data center has up to 10,000 computers
  - Data center at The Dalles, Oregon

Introduction to Question Answering

School of Information Studies
Syracuse University

Question Answering (QA)

IR assumes that the user wants an entire document, while QA assumes that the user wants a **small, focused text result that answers a question,** possibly with some justification to give credibility.

QA systems must apply **more NLP analysis** to text in order to find the answers to questions.

- Answers may not only be phrased with different vocabulary, but they may be implied by other statements or facts.

The document collections of QA systems range in size from the Web to a targeted collection of documents such as a company's product documents.

Question Answering Example

Example where finding the answer to the question (and its justification) requires text processing

Question: *What year did Marco Polo go to Asia?*

Answer is found in this text:
Marco Polo divulged the truth after returning in 1292 from his travels, which included several months on Sumatra.

To answer the question:

- Resolve "his travels" to mean Marco Polo's travels
- World knowledge that Sumatra is in Asia
- "travels" is the noun corresponding to the verb "travel," which can be an instance of the verb "go"
- "returning" from travels is part of the travels and likely occurred in the same year
- So, the answer is "1292"

Typical Traditional QA System

Traditionally, QA systems applied a two-step strategy.

- They assumed that the document collection of a QA system is too large to apply the (time-consuming) NLP processing to all the documents.
- First, use IR to retrieve a set of relevant documents.
- Then process those documents with NLP techniques.
- Identify answers in the documents.

Factoid Questions (From TREC)

Where is Belize located?

What type of bridge is the Golden Gate Bridge?

What is the population of the Bahamas?

How far away is the moon?

What is Francis Scott Key best known for?

What state has the most Indians?

Who invented the paper clip?

How many dogs pull a sled in the Iditarod?

Where did boccie originate?

Name a flying mammal?

How many hexagons are on a soccer ball?

Who is the leader of India?

TREC Question Answering Track

Goal: encourage research into systems that return actual answers

- From 1999–2004 in various forms

Questions were short and fact-based

- From Encarta and Excite search logs

Extract or construct answer from set of documents

For each question

- Supply up to five ranked answer submissions
- With the most likely answer ranked first
- Answer strings were evaluated by NIST's human assessors for correctness

Evaluation: mean-reciprocal rank (MRR)—score given is the reciprocal of the rank of the first correct answer

Evaluation of Answers

How much of an answer is enough?

Q: *Where will the 2002 winter Olympics be held?*

Text: *The 2002 Winter Olympics will be held in beautiful Salt Lake City, Utah.*

- A1: beautiful Salt Lake City, Utah
- A2: Salt Lake City, Utah
- A3: Salt Lake City
- A4: Salt Lake
- A5: Utah

Systems were required to give one of these focused answers to demonstrate their understanding.

- The questioners would probably prefer an entire sentence that established an authority for the answer, perhaps with the answer text highlighted.

Question Answering Techniques

School of Information Studies
Syracuse University

Question Classification

Questions are first analyzed for their type in a question ontology.

- Usually hand-built, but sometimes automatically learned from hand annotated questions

Factoid questions are further classified by an answer-type taxonomy.

- Includes types of entities to be found as an answer
  - "Who founded Virgin Airlines?" Expected type: Person
  - "What Canadian city has the largest population?" City
    - Also location, time, quantities, etc. (Where? How long? How much?)
  - Or, more specifically, for flower, human description, human group, human individual, etc.
    - See Figure 28.4 in J&M, 3rd edition (online)

Question Classification

Question analysis can identify other question types.

- Definition questions
- Reasons (to answer why questions)
- Special forms such as birth and death dates
  - *"When did George Washington die?"*

Answer Processing

Approaches typically have a suite of answer-processing modules that use different techniques to find the answer.

Some are specific to particular question types and expected answer types.

For questions that are looking for particular entity types, find sentences with those types.

- For example, if a question is looking for person, return sentences with named entities of type person.
- Rank those sentences according to the presence of question words.

  Q: *"Who is the prime minister of India?"*

  Answer sentence: *"Manmohan Singh, Prime Minister of India, has told leaders…"*

Specialty Answer Processing

Some question types, such as those querying when someone is born or has died, are often answered with a fairly small set of sentence constructions.

A special purpose module may be built that contains a number of the patterns of those sentences.

- Often hand-built regular expression patterns
- Define patterns of sentences to answer the question
- Use parsed sentences to match those patterns
  - Patterns include subject/verb forms
  - Appositions
  - Known relations for that type of question
- Find additional forms by using bootstrapping similar to relation extraction

Answer Patterns

Query reformulation

- Rephrase query as a partial answer sentence “Virgin Airlines was founded by X” and try to find semantically similar sentences to this particular syntactic pattern.

The Web approach

- Since there are potentially millions of documents on the Web that could answer this question, assume that there is one that answers the question in the same form and vocabulary as the question.
- Search the Web to find the answer in the right form.

Deeper Semantic Reasoning

Query reformulation

- Rephrase a query as a partial semantic relation
  Founded (X, Virgin Airlines)
  And try to find or infer instances of this relation from the text

Semantic processing of the text

- May have relation extraction
- May have parsing and semantic role labeling, using the semantic roles as relations (and identifying semantically similar verbs)
  "Richard Branson formed Virgin Atlantic Airways in 1984, launched Virgin Mobile in 1999, …"
- Will also need to use knowledge sources, such as gazetteers, and some relations such as "near"
- Can use other forms of real-world knowledge, such as from Wikipedia

Selecting the Final Answer(s)

In approaches with multiple answer modules, the results must be combined

- Answer ranking
- Confidence levels

Often difficult to produce comparable confidence scores from different modules

The Watson Question Answering System

School of Information Studies
Syracuse University

IBM's Watson

Deep QA project, headed by David Ferrucci*

- Similar system architecture to typical systems but bigger and faster

Watson interface allowed the Deep QA system to be used to play Jeopardy by selecting questions, deciding when to "ring in" with a confident answer and how to bet

**Ferrucci: Introduction to "This is Watson," IBM Journal of Research, 2012*

IBM's Watson

Real-time challenge

- Need answers in under three seconds
- Optimized hardware, based on IBM Power 7 and software
- Distributed parallel computing for answer modules

Incorporated NLP, machine learning, knowledge representation

- Pre-processed text documents and stored semantic representations for later reasoning
- Instead of IR approach to retrieve documents, used this pre-computed knowledge store with NLP techniques

Question categories in *Jeopardy* add to expected answer types

- Some are entity types
  - Vice presidents, Canadian cities, characters in classic literature
- Some are more complex
  - Rhyme time, starts with "A"

Jeopardy Example Topics and Questions

Recent history
President under whom the United States gave full recognition to Communist China
(Answer: Jimmy Carter)

Pop music
Their grandfather had a number-1 record in 1935; their father, number 1s in 1958 and 1961; and they hit number 1 in 1990
(Answer: Gunnar & Matthew Nelson)

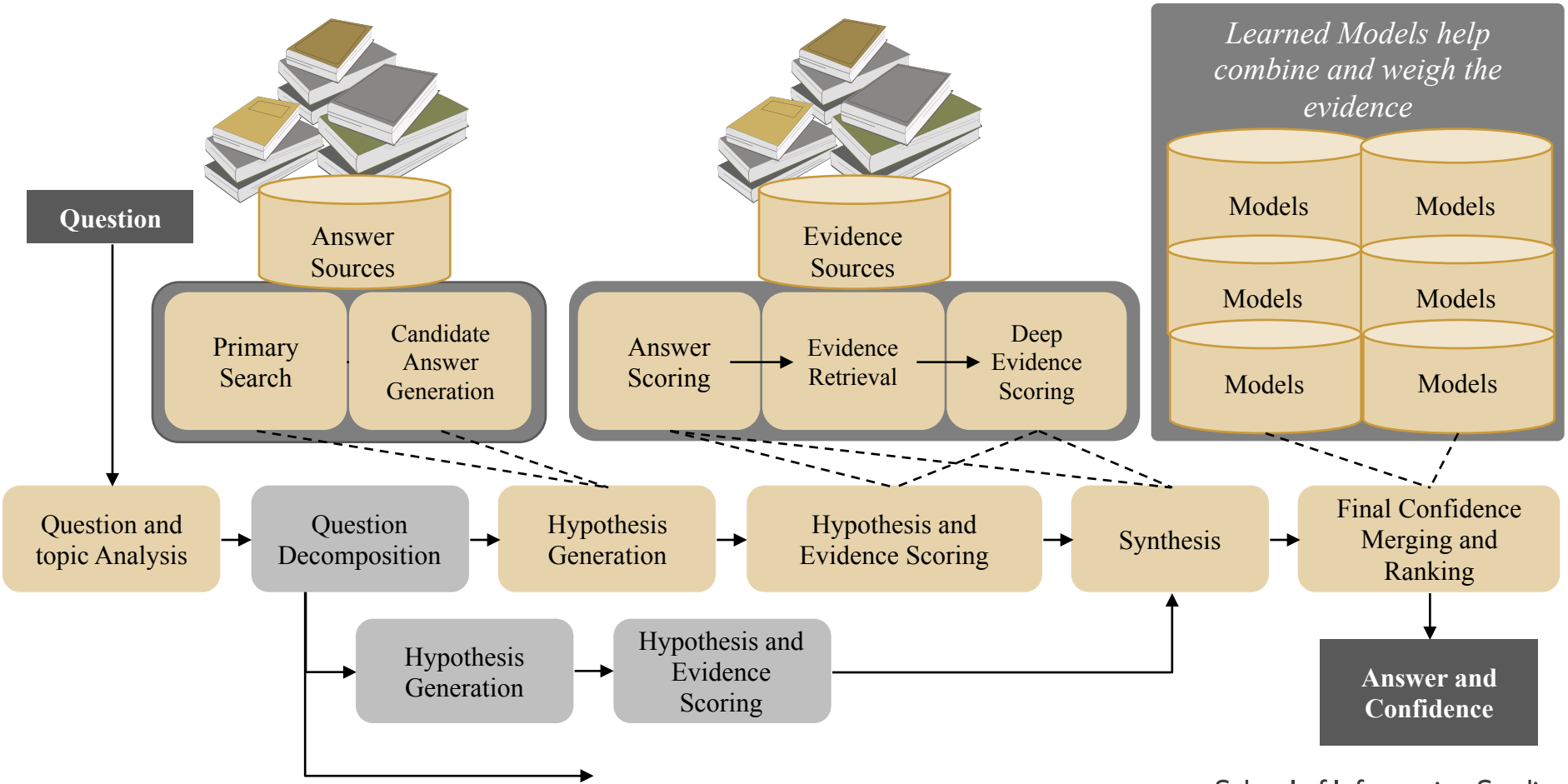
Before and after
The *Jerry Maguire* star who automatically maintains your vehicle's speed
(Answer: Tom Cruise control)

Art history
Unique quality of "First Communion of Anemic Young Girls in the Snow" shown at the 1883 Arts Incoherents Exhibit
(Answer: all white)

Common bonds
Feet, eyebrows, and McDonald's
(Answer: arches)

Language
The lead singer of the band Dengue Fever is from this country and often sings in Khmer
(Answer: Cambodia)

Watson Deep QA Architecture



Watson's Data

Not connected to the Internet during the game

- Stored approximately 4 terabytes of data
- **Creation of knowledge store one of major contributions**

"Deep analytics" of text to extract relations and facts

- Knowledge representation, primarily in the form of relations, both from relation extraction and semantic role labeling

Processed 200 million pages of structured and unstructured data, including the entire Wikipedia

- Structured sources like gazetteers
- But only 2% of questions answered directly from structured data; all others used, at least partly, information from text

Sources included encyclopedias, dictionaries, thesauruses, newswire articles, and literary works

Watson's Steps

Question analysis: type of question, what the question is asking for

Hypothesis generation by "thousands" of tasks

- Quantity over accuracy

Hypothesis and evidence scoring

- Collects positive and negative evidence from passages
- Uses relations between concepts learned from text
  - Books have authors, authors create characters

Final merging and ranking

- Uses past experience to weigh types of evidence for different questions
- Uses confidence of top answer to decide whether to "buzz in"
  - Wagering strategy
- **Deciding and merging confidence another major contribution**

Questions Requiring Complex Answers

School of Information Studies
Syracuse University

Another Type of Information Search: Complex Questions

Many questions go beyond asking for simple facts.

- Comparison questions
- Questions that involve deeper understanding of issues

These questions cannot typically be answered by a simple phrase or sentence.

Approaches focus on identifying an answer passage.

Example Passage Retrieval System

CNLP at Syracuse built a domain-specific QA system to support online aeronautics class

- National Aeronautic & Space Agency funding

Collection of resources from professors of the class

- Textbooks, technical papers/reports, websites
- Pre-selected for relevance and pedagogical value

Example Passage Retrieval System

Community of users focused on specific tasks

- Undergraduate students from two universities majoring in aerospace engineering
- Students using system for a course can ask questions while working in teams or on own

Where QA system must function:

- In real time, not batch mode
- On real users' real-world questions
- With real, not surrogate assessments of relevance

Sample Questions From Real Users

- *How difficult is it to mold and shape graphite epoxies compared with alloys or ceramics that may be used for thermal protective applications?*
- *How dose the Shuttle fly?*
- *Do welding sites yield any structural weaknesses that could be a threat for failure?*
- *Are thermal protection systems (TPSs) of spacecrafts commonly composed of one panel or a collection of smaller tiles?*
- *How can the aerogels be used in insulation of holes in TPSs?*

Two-stage QA model

- **Passage retrieval using expanded query representation**
- **Selection of answer: providing passages based on generic and specialized entities and relations**

Conversational Agents

School of Information Studies
Syracuse University

Conversational Agents

Conversational agents are the general class of dialogue systems that communicate with people with a natural language interface.

Task-oriented dialogue agents are designed to help users achieve a particular task or tasks and to have short conversations.

- Digital assistants such as Siri, Cortana, Alexa, etc.

Chatbots are designed to have more extended conversations with humans, mimicking human/human interactions.

- Often designed for Turing test of passing for human

Informal terminology often uses the term "chatbot" for both of these.

Chatbots

Rule-based chatbots include the most famous one, the original ELIZA system from Weizenbaum in 1966 that simulated a Rogerian psychologist.

- User: *Everybody laughed at me.*
- System: *Why do you think everybody laughed at you?*

- Rules pick out particular keywords (you, me) and have a set of responses that transforms the user's statements.

More recent corpus-based chatbots use logs of human conversations to map user utterances into responses.

The goal of modern chatbots is often entertainment value.

Frame-Based Dialogue Agents

Current task-based dialogue systems are often based on a domain ontology, for one or more domains, such as travel, calendar, dining, etc.

- Based on **user intent**
- System has one or more **frames** with **slots** for specific pieces of information to obtain from the user
- **Question** for each slot

Slot	Type
ORIGIN CITY	city
DESTINATION CITY	city
DEPARTURE TIME	time
DEPARTURE DATE	date
ARRIVAL TIME	time
ARRIVAL DATE	date

Slot	Question
ORIGIN CITY	"From what city are you leaving?"
DESTINATION CITY	"Where are you going?"
DEPARTURE TIME	"When would you like to leave?"
ARRIVAL TIME	"When do you want to arrive?"

Operation of Dialogue Agents

System carries on a mixed initiative dialogue with user

- System generally asks questions to fill slots but must first detect user domain and intent
- User may also override conversation to steer into other channels

System must use natural language understanding to process user utterances

- Detect domain shift and intent
  - *I'd like to book a rental car when I arrive at the airport.*
- Flexibly fill slots by detecting possible multiple pieces of information or no information from the user
  - *I want a flight from San Francisco to Denver one way leaving after 5 p.m. on Tuesday.*
- Deal with speech recognition errors or other situations not in its frames

Dialogue Agent Systems

Commercial systems allow developers to implement frame-based systems by:

- Defining domains and intents
- Defining frames, slots, and question responses

VoiceXML is a format for specifying such a system, although simpler than commercial systems.

- Adds a "no-input" response: *I'm sorry, I didn't hear you.*
- And a "no-match" response: *I'm sorry, I didn't understand you.*

Digital assistants such as Apple's Siri, Amazon's Alexa, and the Google assistant are frame-based.

- Adds more sophistication in initial conversations, including some question answering capabilities
- Some allow additional user domains to be given

Dialogue System Design

Dialogue systems are a kind of human–computer interaction, and general HCI (human–computer interface) design principles apply.

- Study the user and task.
- Build simulations and prototypes.
  - Wizard-of-Oz system has a human simulate the system
- Iteratively test the design on users.