University of Illinois at Urbana-Champaign
Dept. of Electrical and Computer Engineering

ECE 101: Exploring Digital Information Technologies for Non-Engineers

Search Engines

The Turing test: A Measure of Artificial Intelligence

- Created by Alan Turing
- A method for evaluating a machine's ability to simulate human behavior and intelligence
- The Imitation Game: A game between a human, a machine, and an interrogator, where the interrogator tries to determine which of the other two is the human
- If the interrogator is unable to determine which participant is human, the machine is considered to have passed the test
- A computer can be said to possess artificial intelligence if it can mimic human responses under specific conditions.

Until Recently, Computers Could Not Perform Recognition

What question would you ask?

Maybe: How many humans are in the picture?

A decade ago, a **computer** would **not** be **able to answer**.

Neither Could Computers Read Oddly Written Text

How about a captcha?

Again, **computers could not identify the characters**.

That's why captchas were useful in preventing bots from creating accounts and doing other things.

Intelligence: Just what is It, Anyway?

The notion of intelligence is a subject of much investigation.

What does it take to make machines “intelligent?”

- **Is it the ability to do a specific task exceptionally well?**
- **Is it the ability to adapt to the task and the environment?**
- **Is it the ability to perform a wide range of cognitive tasks?**

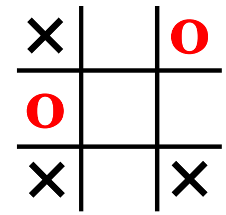
Artificial Intelligence vs. Artificial General Intelligence?

- **Artificial General Intelligence**—can match or surpass human abilities. (Don't mix up AGI with Generative AI)
- **Artificial Intelligence**—can do a particular task as well as human (Often "machine learning" is mislabeled as "AI")
  - can see and identify objects
  - can understand and respond to human language
  - can learn from new information and experience
  - can make detailed recommendations to users and experts
  - can act independently, replacing the need for human intelligence or intervention (example: self-driving car).

Search is Fundamental to Machine Intelligence

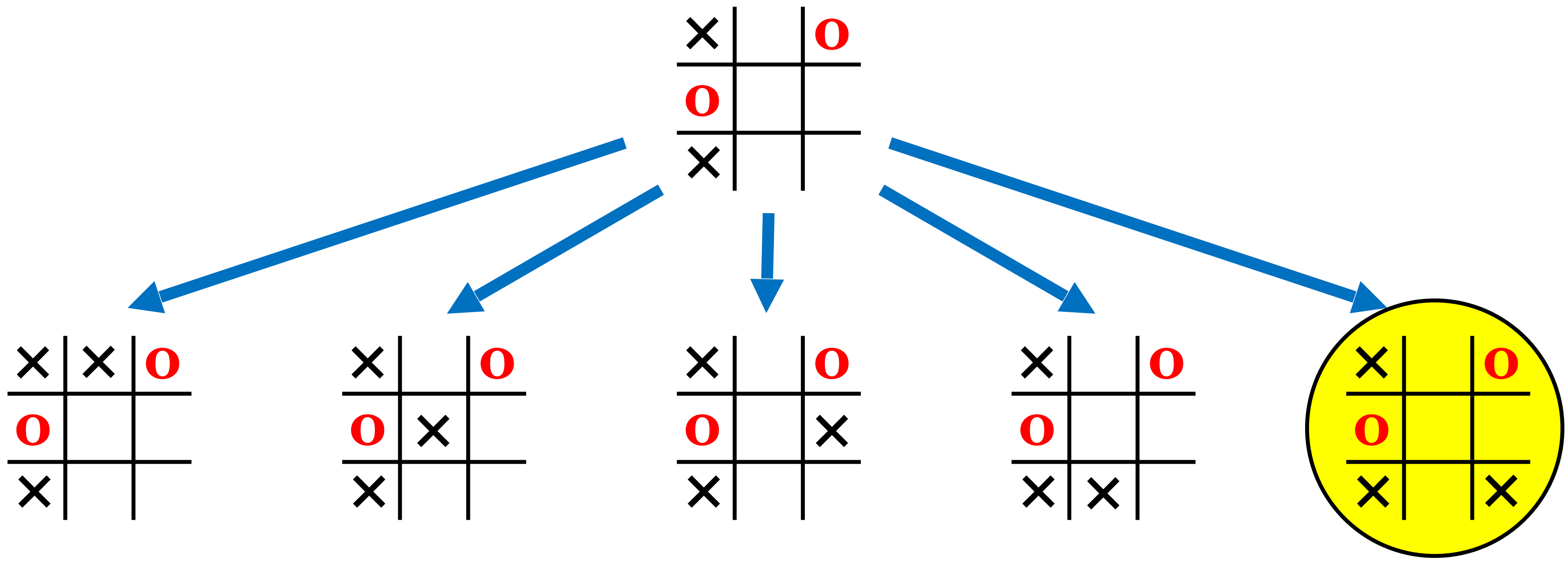
We claim that
search is fundamental to intelligence.

Say for example:
- Playing tic-tac-toe or
- Identifying animals in photos.



Searching for a Move in Tic-Tac-Toe

What is the next best move for ✕ ?



Object Recognition an Important Part of Computer Vision

What do you see in the picture?

Identifying animals,
- even if it's just a dog or a cat,
- seems like an "intelligent" action.

In fact,
- **object recognition** has been
- **an important part of computer vision**
- for decades.

Best Recognizers Now Use Search, Appear Intelligent

Recently, though,
- the **approach taken** by the best systems
- has been **more like a search problem**.

And computers are good at those.

If machines
- **can solve search** problems **quickly**,
- they **can appear intelligent**.

A More General Example: Web Search

Here's another example:

There's so much content scattered over the edge of the Internet.

Given a topic,

- how can I find
- the most relevant and interesting pages?

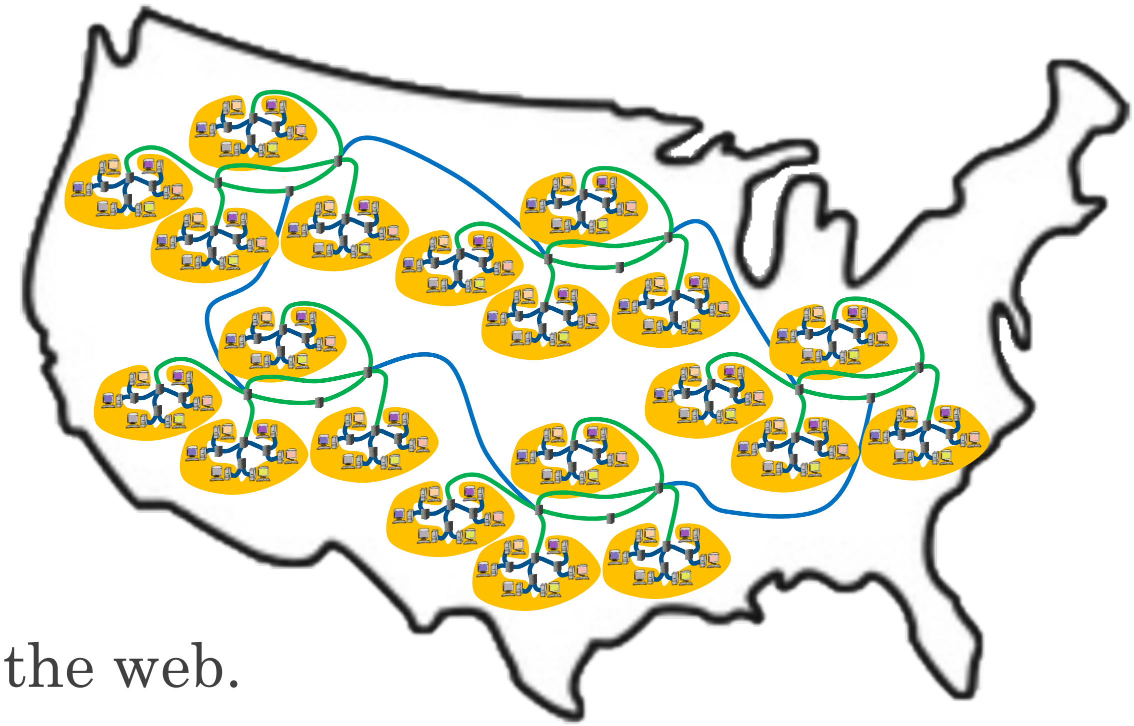
Search Engines: Find Interesting Content at the Edge

Content is scattered over edge of Internet (orange).

More and more content as Internet grows.

The **search engine problem: how to find documents of interest?**

Specifically, find words/concepts in a set of documents scattered over the web.



Searching Example: Find the URL in Milliseconds

Example: I want the ECE 101 web page.

But where is it?

Need to **search** all web pages
- **and find the right one**.
- Should be possible: every page has a URL.

Want **results in milliseconds**:
- studies suggest going **beyond 400ms**
- starts to **lose customers**.

How Does Web Search Operate?

Steps in an Internet search service:

1. User **searches with phrase P** ("ECE 101")
2. Search engine **finds**
   - a large **set of** scattered **documents D**
   - **that contain P** or are relevant to **P**:
   - $\mathbf{D = \{d_1, d_2, \ldots, d_{203}\}}$.

How Does Web Search Operate?

3. Search engine **re-orders D by "relevance" to obtain R**:
   - **R = {d_{17}, d_{185}, d_{23}, d_{29}, …, d_{12}}**
   - "relevance" could depend on the searcher
     - History of Nike (history professor) -> Greek goddess
     - History of Nike (political scientist) -> Nike missile
     - History of Nike (CEO of company) -> Nike shoes

4. **Show results** of search **quickly**.

Search is a Five Step Process

How does the search engine find the information?

1. **Crawl the web** to gather billions of documents (text, images, videos).
2. Organize the documents for fast searching,
   - called **indexing the documents** (like the index in a textbook)
   - **to create** a set **S**.
3. **Order** documents in **S by** decreasing **reputation**.

These three occur **before anyone does a search**.

Search is a Five Step Process

Two more steps, using the search phrase:

4. Use phrase **P** to **filter documents in S** in order **to find relevant set R**.
5. Do another round of **reorder**ing
   - **based on knowledge of user**
   - (search history, YouTube preferences, travel, purchases, and so on)
   - **to create** the **list L** for display (ads go in front!).

First Step of Search: First Crawl the Web Graph

As you know,
- a **URL** lets us access a web page;
- these are the **nodes** in the web graph.

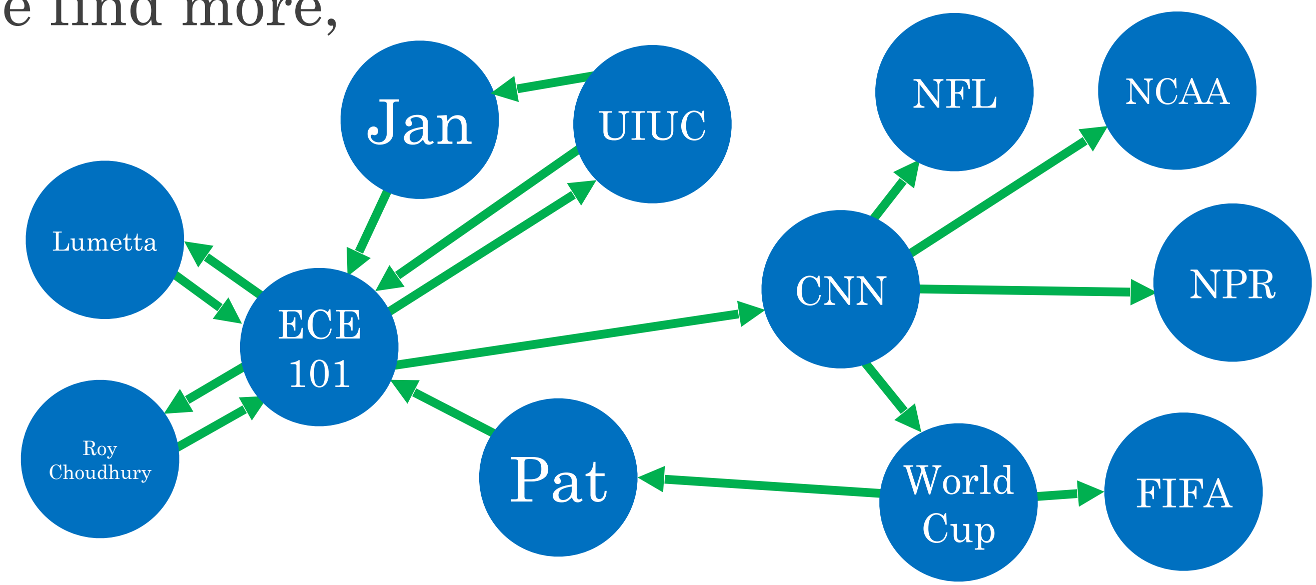
In each web page are more URLs:
- **links to other pages**, images, videos, and so forth;
- these **are the edges** in the web graph
- (directed: point from one node/URL to another).

Web Crawler Tries to Find All Content in the Web

Starting with the class page,
- we find several links.
- In those pages, we find more,
- And more.
- And more.

Crawler keeps moving.

Tries to find everything.



Second Step of Search: Index the Documents

The gathered documents must be indexed.

Documents need to be easy to find quickly.

How can one accomplish this task?

What if we create a separate folder
- for each word (say in English), then
- place a document in the folder if it contains the folder's word?

Relevant Information Must be Found Quickly

We still need to find the folders quickly!

For example … find the word "precise" in a dictionary.

No, not an online dictionary / search engine.

This kind…

The words are all sorted alphabetically.

For most of you, it's been a while, right?

A Challenging Use of Old Technology…?

How can you find "precise" quickly?

Some Constraints on Our Answers

Let's limit our approach to something that works well for any word.

In other words, we don't reason
- that "p" should be about 2/3 through the dictionary
- nor make use of a "P" tab in the side (many thicker dictionaries provide them).

Solution? Divide into Halves Repeatedly

Then, ideally,
- Open to exactly the middle.
- If "precise" comes before
  - the words on that page,
  - "precise" must be in the first half.
- If "precise" comes after
  - the words on that page,
  - "precise" must be in the second half.

Start over with half of the book.

Repeat until we find the right page.

Binary Search Finds any Word Fairly Quickly

After 10 steps,
- a dictionary with **22,000** pages,
- such as the Oxford English Dictionary,
- is **reduced by 2^{10} = 1024**, leaving only
- about **21.5** pages to search.

After another **7** steps,
- we have only about **1/6th** of a page
- in which to find our word.

That **approach** is **called binary search** (because we divide each part into two parts).

Third Step of Search: Sort the Documents by …

We have billions of documents from the Web.

Which ones are more important, relevant, and/or accurate?

Which ones are less so?

How can we possibly decide?

Time to sort … by reputation

Maybe we can let the web graph decide for us?

A Page is “Good” if Others Point to It

Idea: if a page is important, other pages will link to it.

For example, links to reference articles: “If you want to understand binary search, you can read about it on Wikipedia: https://en.wikipedia.org/wiki/Binary_search_algorithm)”

Or health information: C-U Public Health Department’s COVID vaccination information: https://www.c-uphd.org/covid-vaccinations.html

Or recommendations: “I took this great class last semester! Check it out: https://courses.grainger.illinois.edu/ece101/fa2023/”

Importance: Number of Incoming Edges in the Web Graph

So we can assign **"importance"**
- based on **how many other pages**
- **link to a page**.

But that **may not be enough**.

For example,
- if the New York Times has a pointer to a page,
- that fact may be (slightly) more important
- than the link to the same page from mymonekytyping.com.

Links from Important Pages are More Important

When we count incoming edges,

- we **also want to count**
- the "**importance**" **of** the **pages**
- **that link to a page.**

But we should also count

- the "importance" of the pages
- that link to those pages.
- And so on.

Page Rank Captures "Importance" in a Directed Graph

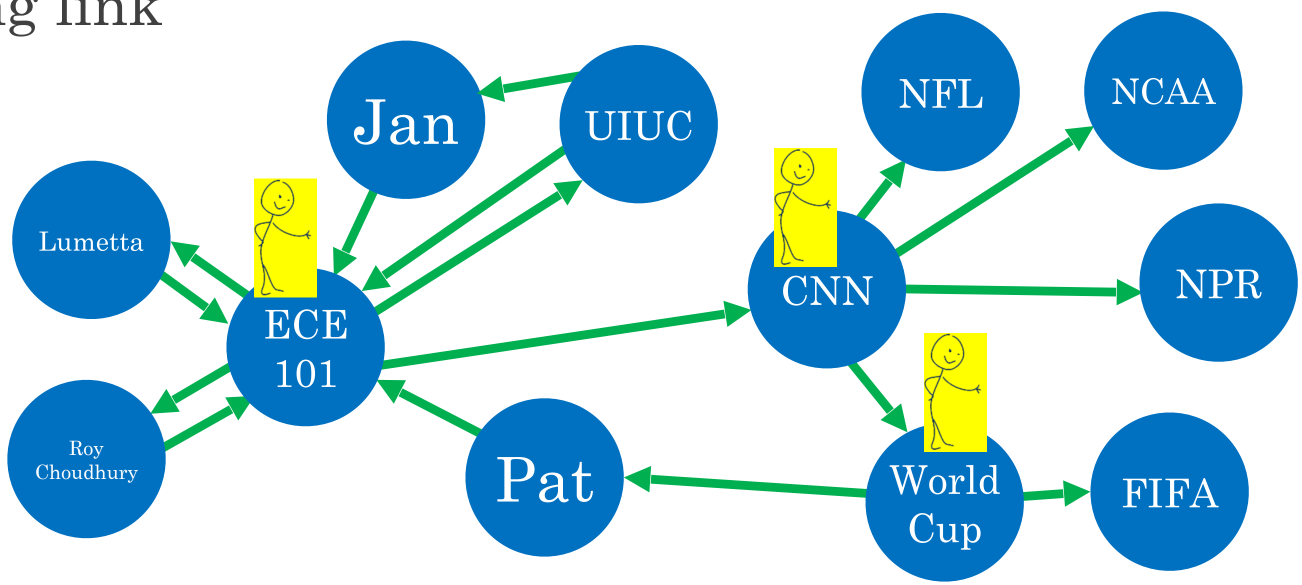
Google developed an idea
- known as **page rank**
- **to capture this idea**.

The same approach
- is now popular
- for many purposes
- in graph problems.

Page Rank Intuition: People Walking at Random (Video)

Imagine a person at every node (some shown).

Each person
- chooses an outgoing link at random
- (equal probability, independently of past/future decisions)
- and walks to another node.



Page Rank: Expected Number of "People" At a Node

One issue:

- some pages include no URLs.
- Anyone at such a node
  - can 'start over'
  - by choosing a new node at random
  - instead of going down a link.

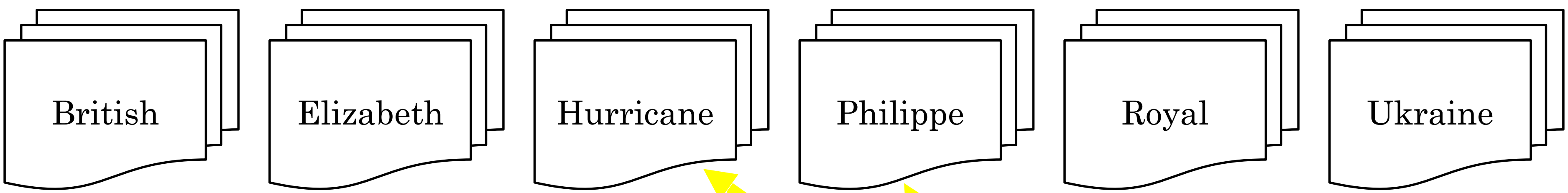
Repeat the process many times, then

- count how many people are at a node
- to find the node's "importance" (rank).

Search Indexed Documents for a Given Phrase

Fourth step: filter the documents with phrase P.

Let's imagine we have folders indexed by words…



We search for "Hurricane Philippe"

What should we do?

Intersect contents of these two folders.

Fifth Step of Search: Apply Knowledge of User

Reorder search results based on knowledge of user.

This part is proprietary:

- how it's done is specific to the individual company,
- as is the information on which it is based, and
- they don't tell anyone the details.

Doing a better job than other companies keeps

- advertising revenue coming in and
- customers coming back.

Some User-Based Changes Can be Deduced

Nevertheless, we can observe some of the choices made directly.

Importance increases based on location and search history:

- For example, if searcher comes from Illinois.edu, all web pages at UIUC are boosted in rank.
- If you view the ECE 101 web page two or three times a week, all links from that page (and those pages, and so forth) are boosted in rank.

Search Engine is Not Using Your Local Data

A student wondered,

"**What happens if you clear your browsing history?**"

Search engine companies

- track your history separately.
- **Your** local **copy is not used**.

The **companies** also **use information** about

- **your IP address** (where are you?),
- **your browser**, and
- more or less **anything else that they can deduce** by having figured out who you are and consulting their records on your preferences.

Search is Curiously "Smart"

For example, **when I look for "ECE 101"**

- **I get our class page**.
- **Even if** I use **"private"** browsing (on Google, Bing, or Yahoo).
- **Even if** I **use Duck Duck Go**, which says it will not track me (but what about my IP address in Champaign)?

The only search engine that doesn't give me the class page is Baidu, for which results in English don't rank highly…

Tracking Can be Beneficial

Sometimes tracking is really useful.
When?

- Weather tracking - a tornado warning
- Personalized experience
- Find deals you like
- You don't mind providing info …

Tracking Can Also Lead to Confusion and Embarrassment

You probably share an IP address with your family or roommates,
so some information may be mis-attributed to you,
and some ads for you may show up in their browser.

Lack of Privacy!!

Europe Has Led the Way in Privacy Regulation

Thanks to European privacy laws

- (General Data Protection Regulations, or GDPR),
- **companies are being forced**
- **to make** the **data** that they have collected
- **on a person available to that person**
- (and editable, and subject to deletion).

As of today, the **US has no such laws**,

- but some states have started to formulate them, and
- **some companies do make the same data available** to any of their customers.

All Steps Must be Executed Quickly

All of these must be extraordinarily fast.

Use fast computers working in parallel.

10,000 high-end servers working together.

Put common phrases into their own folders (examples: US elections, cat videos, …)

Filter out irrelevant pages early (auto-/randomly-generated, nonsense, spider traps—people trying to defeat crawling, and so forth).

Many years of research and still an interesting problem…

Form of intelligence?

Terminology You Should Know from These Slides

- search engine
- Web crawling
- indexing
- page rank / reputation
- filter (documents by keyword)
- General Data Protection Regulation, GDPR

Concepts You Should Know from These Slides

- problem solved by search engines
- search engines' need for speed
- steps in Internet search from both user and 'anatomy' (internal) perspectives
- personalization vs. privacy in web search
- status of privacy laws with regard to user tracking
- how page ranks works to compute a page's importance