PURDUE UNIVERSITY

CS47300: Web Information Search and Management

Deep Web

Prof. Chris Clifton

17 November 2017

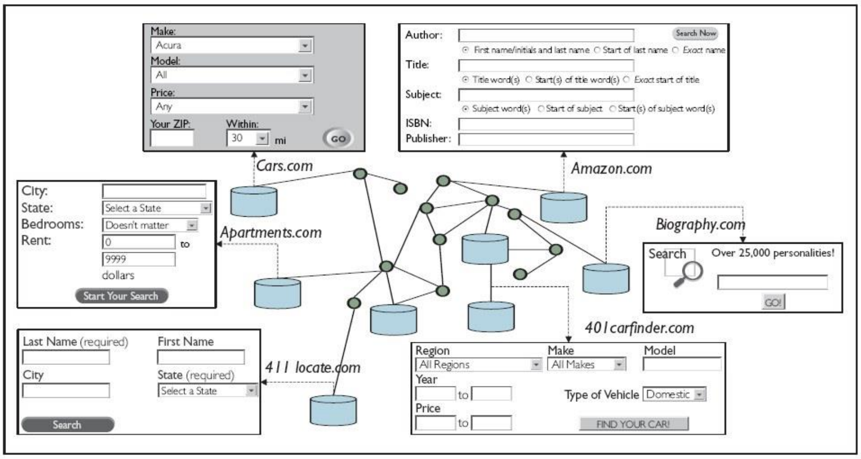
Indiana Center for Database Systems ™

PURDUE UNIVERSITY

Deep Web vs. Dark Web

- Dark Web: Hidden intentionally
  - Largely to support illegal or socially unacceptable activity
  - *But legality and acceptability vary, web is trans-national and trans-cultural*
  - We won't go here…
- Deep Web: Data hidden behind interfaces
  - Federated search is one answer
  - But can we crawl this data?

Conceptual View
(He, Patel, Zhang, Chang '07)



Make: Acura
Model: All
Price: Any
Your ZIP:
Within: 30 mi
GO

Author:
Search Now
● First name/initials and last name ○ Start of last name ○ Exact name
Title:
● Title word(s) ○ Start(s) of title word(s) ○ Exact start of title
Subject:
● Subject word(s) ○ Start of subject ○ Start(s) of subject word(s)
ISBN:
Publisher:

City:
State: Select a State
Bedrooms: Doesn't matter
Rent: 0 to 9999 dollars
Start Your Search

Search
Over 25,000 personalities!
GO!

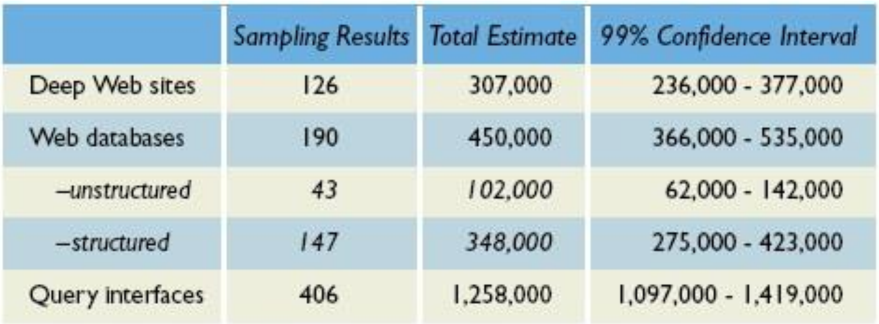
Last Name (required)
First Name
City
State (required) Select a State
Search

Region All Regions
Make All Makes
Model
Year to
Type of Vehicle Domestic
Price to
FIND YOUR CAR!

Does the Deep Web Matter?

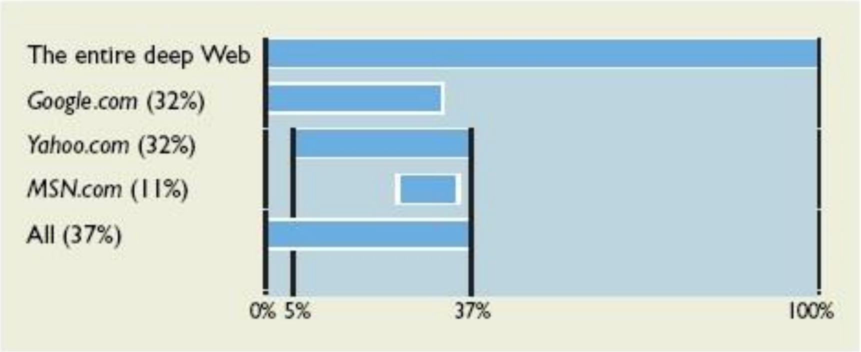
- Where are the entry points?
- What is the scale?
- How "structured" is the data?
- What topics are covered?
- How well do search engines already cover this?
- Wat about existing specialized portals?

Size Estimate of the Deep Web

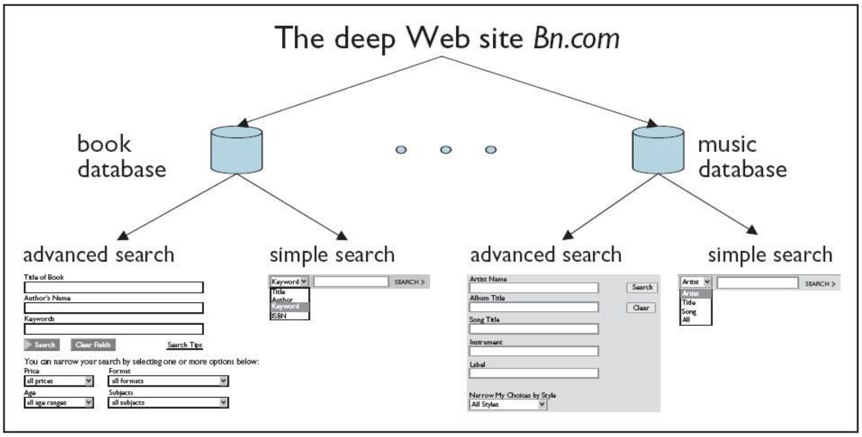


	Sampling Results	Total Estimate	99% Confidence Interval
Deep Web sites	126	307,000	236,000 - 377,000
Web databases	190	450,000	366,000 - 535,000
–unstructured	*43*	*102,000*	62,000 - 142,000
–structured	*147*	*348,000*	275,000 - 423,000
Query interfaces	406	1,258,000	1,097,000 - 1,419,000

Search Engine Coverage



Deep Web Components *(He, Patel, Zhang, Chang '07)*



The deep Web site *Bn.com*

book database

music database

advanced search

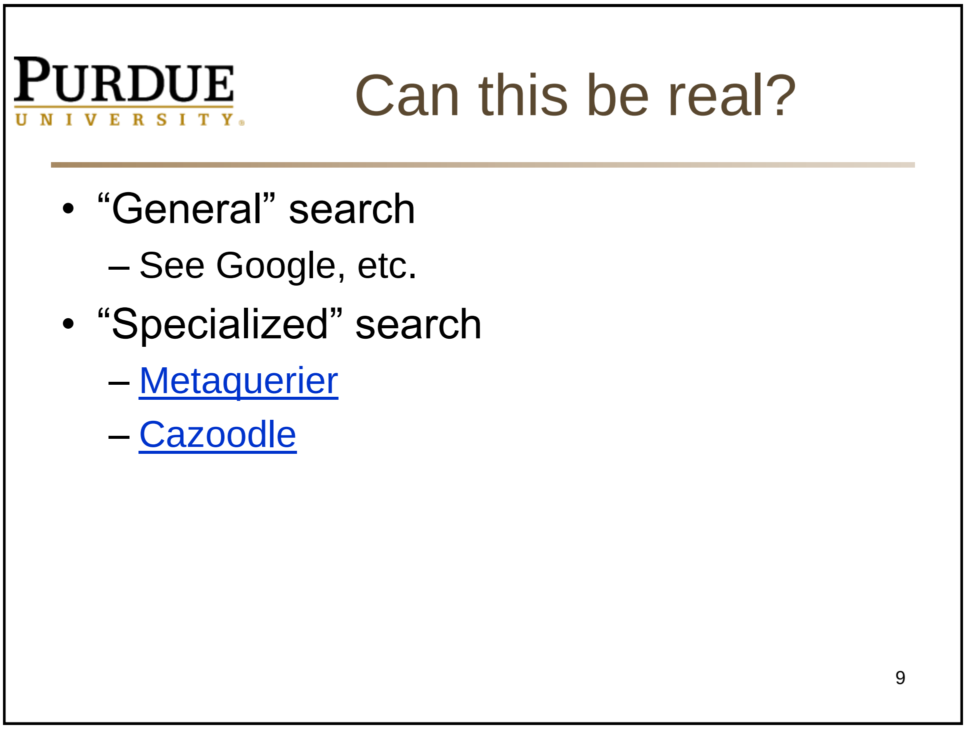
simple search

advanced search

simple search

Challenges

- How do we know what is in a database?
  - Sample queries?
  - Search page
    - Descriptive information
    - Form fields
- How do we query it?
- How do we process results?



Can this be real?

- “General” search
  - See Google, etc.
- “Specialized” search
  - Metaquerier
  - Cazoodle