IL PROBLEMA DELLA RICERCA DI INFORMAZIONI
su Internet e sul web

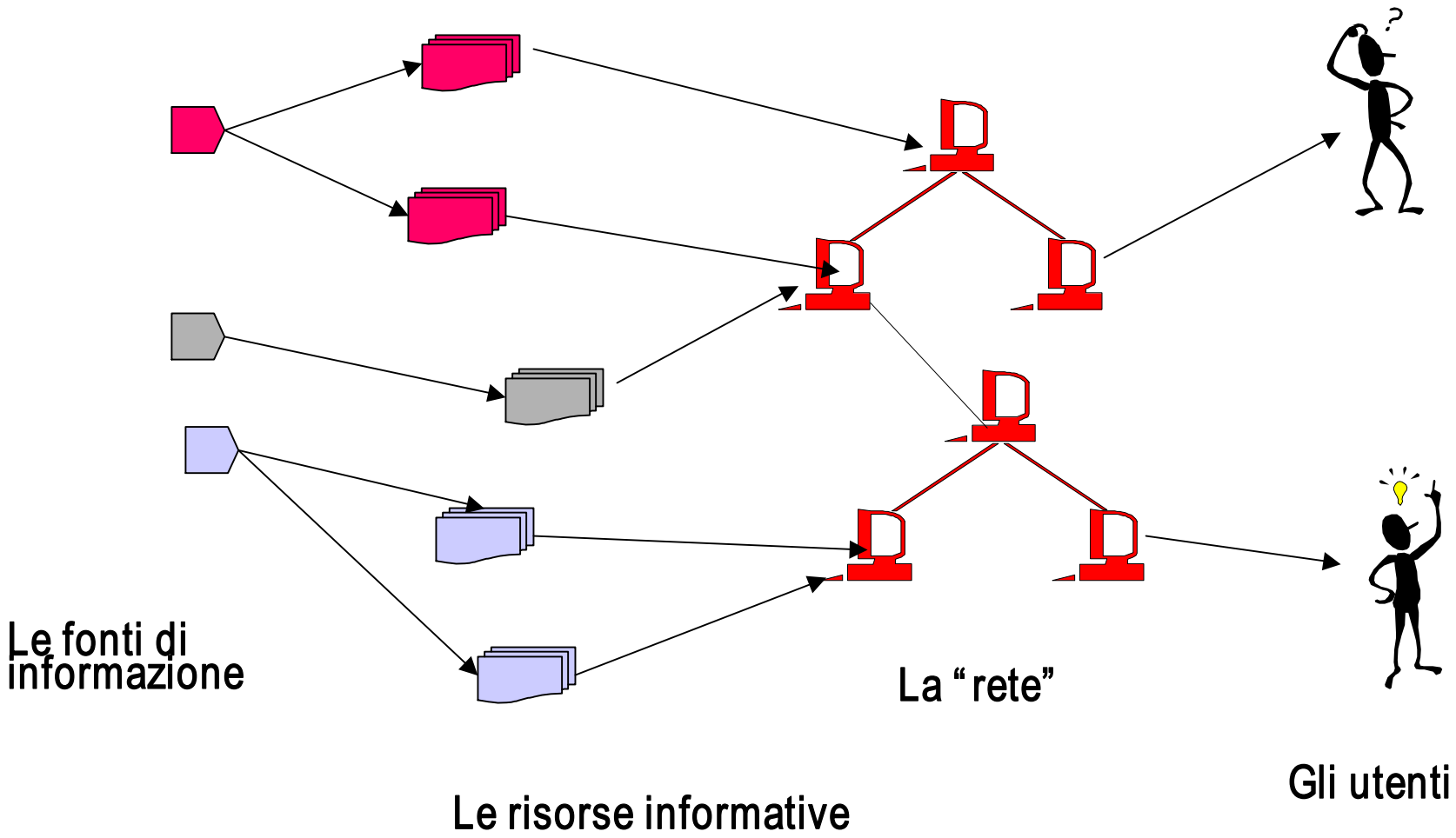
GLI INTERROGATIVI

1. Perché Internet è il più grande contenitore di info del mondo?
2. Perché non è sempre facile reperire informazione utile e affidabile?
3. Quali criteri e strumenti per orientarsi nella ricerca in Internet?

Perché questi interrogativi sono importanti?

1. Perché Internet sta diventando il riferimento fondamentale per la ricerca in ogni campo
2. Perché analoghe problematiche si affrontano in sistemi di informazione di dimensione minore

IL SISTEMA INTERNET COME DEPOSITO DI INFORMAZIONI:UNO SCHEMA



Primo elemento:
CARATTERI DISTINTIVI DELLA RETE INTERNET

- facilità d'accesso (per fonti e utenti)
  - varie modalità (rete dati, telefonica, satellite,…)
  - costi (relativamente) modesti
  - protocolli standard largamente diffusi
- dimensione in continua espansione potenzialmente "illimitata"
- struttura non gerarchica
  - struttura "peer-to-peer"
  - "link"ipertestuali
- ambiente dinamico e "ricco"
  - configurazione e struttura facilmente modificabile
  - supporto multimediale (informazioni "ricche"; elaborazioni)
- varietà dei canali di comunicazione
  - possibile bidirezionalità
  - one to one, one to many, many to many, ecc.

Secondo elemento:
FONTI E RISORSE DI INFORMAZIONE SU INTERNET

- estrema eterogeneità delle fonti
  - istituzioni, aziende, singoli individui, ….
- estrema varietà delle informazioni immesse
  - come contenuti, formati, ….
- assenza di censura/controllo
- facilità di “aggiornamento”
- varie modalità di fornitura
  - es: informazioni protette, a pagamento, oppure libere, etc.
- collegamenti multidimensionali, multilivello, ridondanti

Terzo elemento
UTENTI DELLE INFORMAZIONI SU INTERNET

- estrema eterogeneità dei *fabbisogni informativi*
  - tra utenti diversi
  - per lo stesso utente
- diverse *modalità di accesso*
  - tempi, costi, disponibilità
- numero crescente di *non specialisti*

Quali caratteristiche ha questo sistema, perché è così e in cosa è diverso da altri “depositi” di informazione?

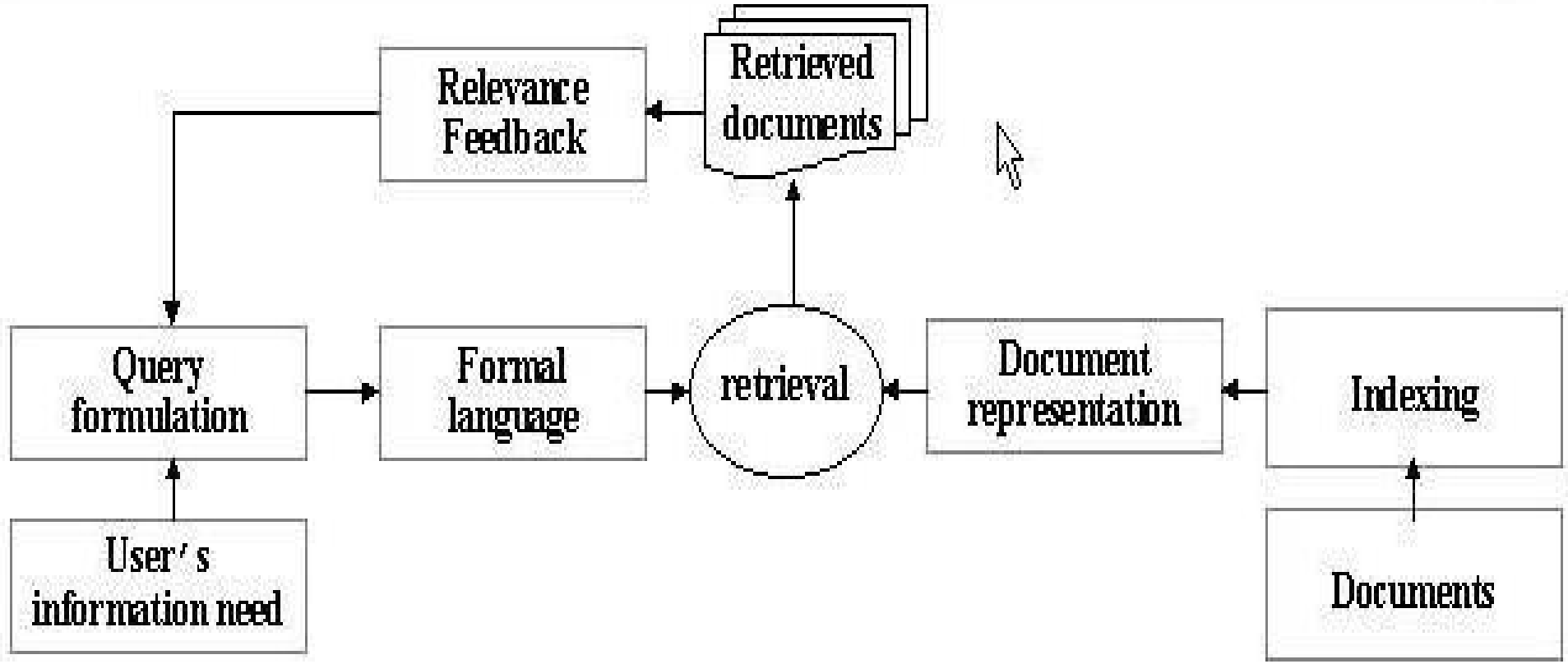
INTERNET COME RISORSA INFORMATIVA: le “parole chiave”

- DIMENSIONE
- VARIETA’ – ETEROGENEITA’
- MULTIMEDIALITA’
- FACILITA’ D’ACCESSO
- DINAMICITA’ (EVOLUZIONE CONTINUA)
- ASSENZA DI CONTROLLO E GERACHIA

Come reperire informazione?

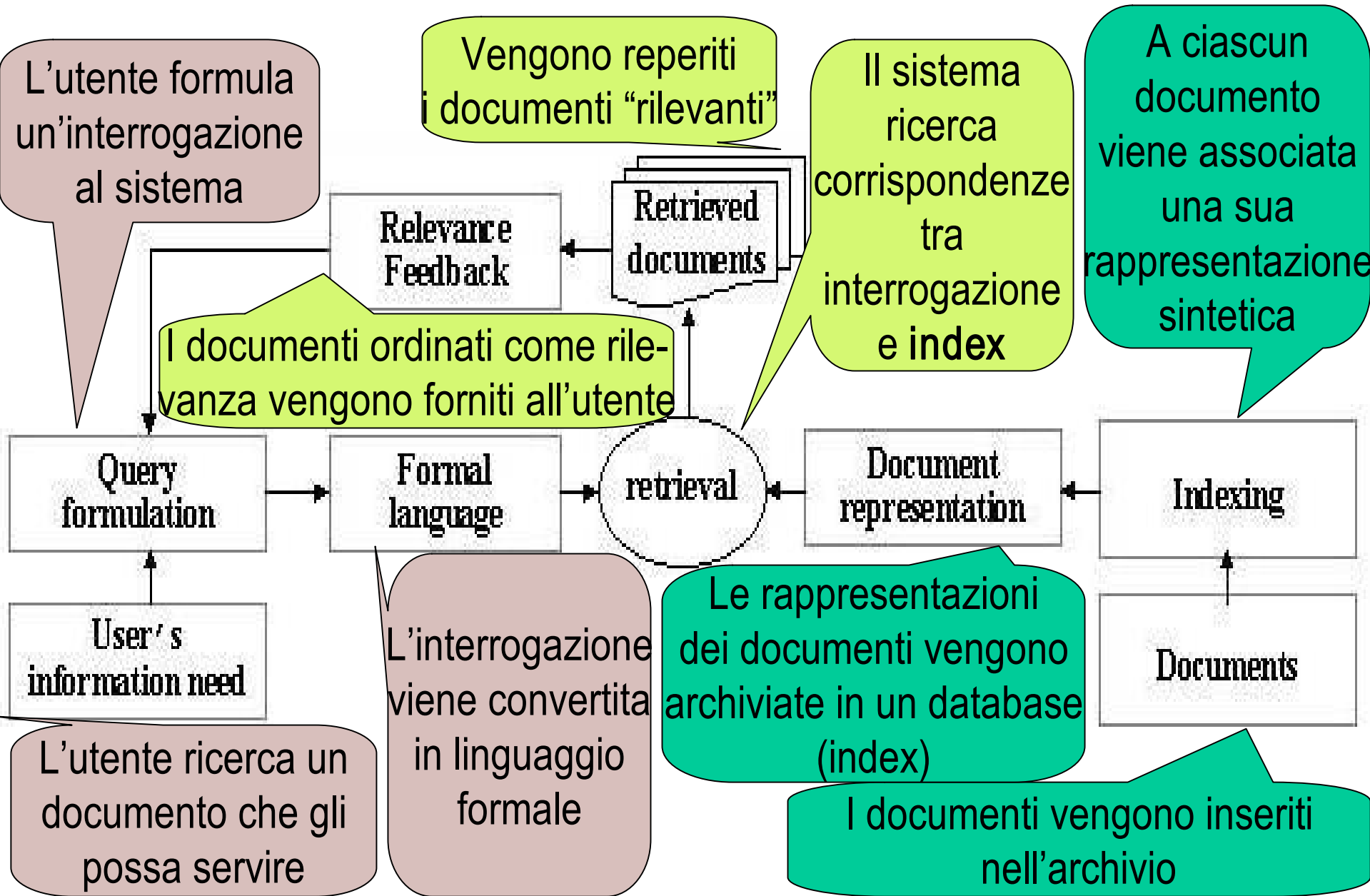
Information Retrieval “classico”

usato nei tradizionali archivi documentali elettronici



Information Retrieval “classico”

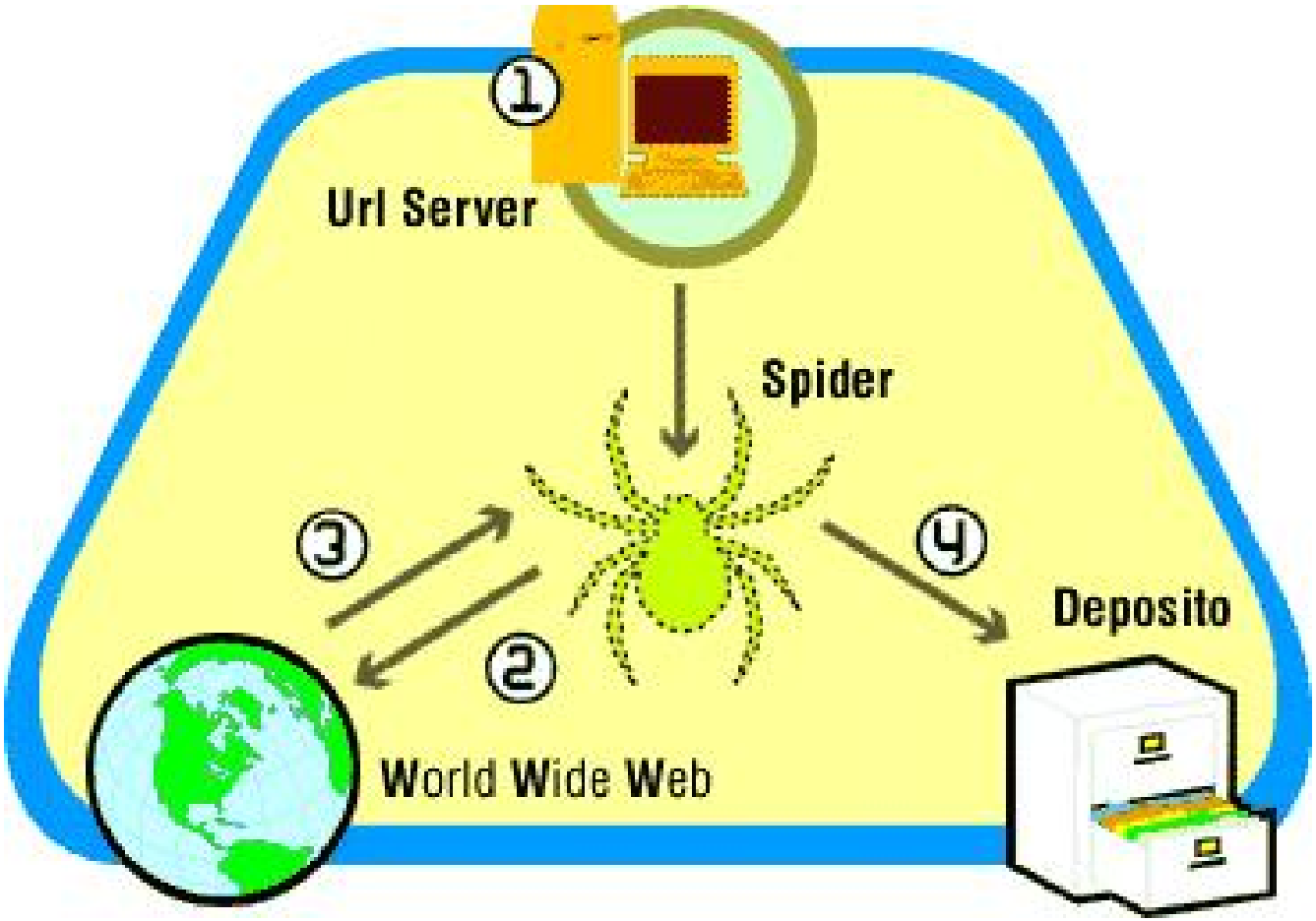
usato nei tradizionali archivi documentali elettronici



Tecniche di indicizzazione (e di retrieval)

- La ricorrenza di ciascuna parola
  - un documento viene rappresentato sulla base del numero di volte che ciascun termine compare
- La distanza tra parole nel testo
- La tecnica vettoriale
- ….

L'APPRPCCIO IR NEL WEB: I MOTORI DI RICERCA



Perché i motori di ricerca non sempre funzionano come sperato?

I problemi dei motori di ricerca (1/5)

DIFFICOLTÀ DI REPERIMENTO DELLE PAGINE

- mancanza di catalogazione
  - struttura non gerarchica
  - elevato dinamismo
  - mancanza di controllo d'accesso
- varietà di formati

QUINDI

COPERTURA INCOMPLETA

NESSUNO > 10-20% di pagine Web coperte!

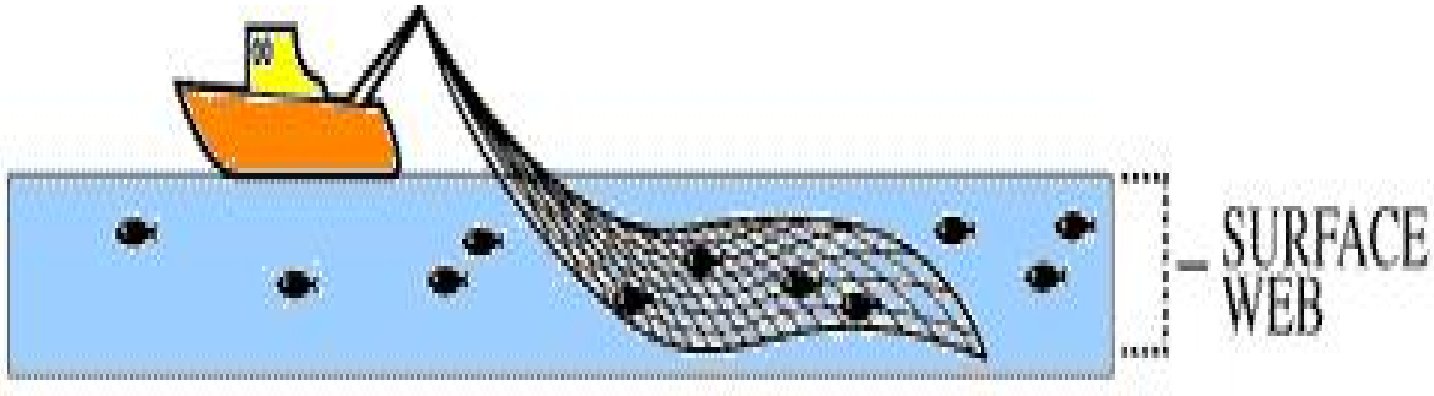
Google 2003: 3 miliardi di pagine (contro circa 20 miliardi stimati)

I problemi dei motori di ricerca (2/5)

- VARIETÀ DI LIVELLI: IL “DEEP WEB”

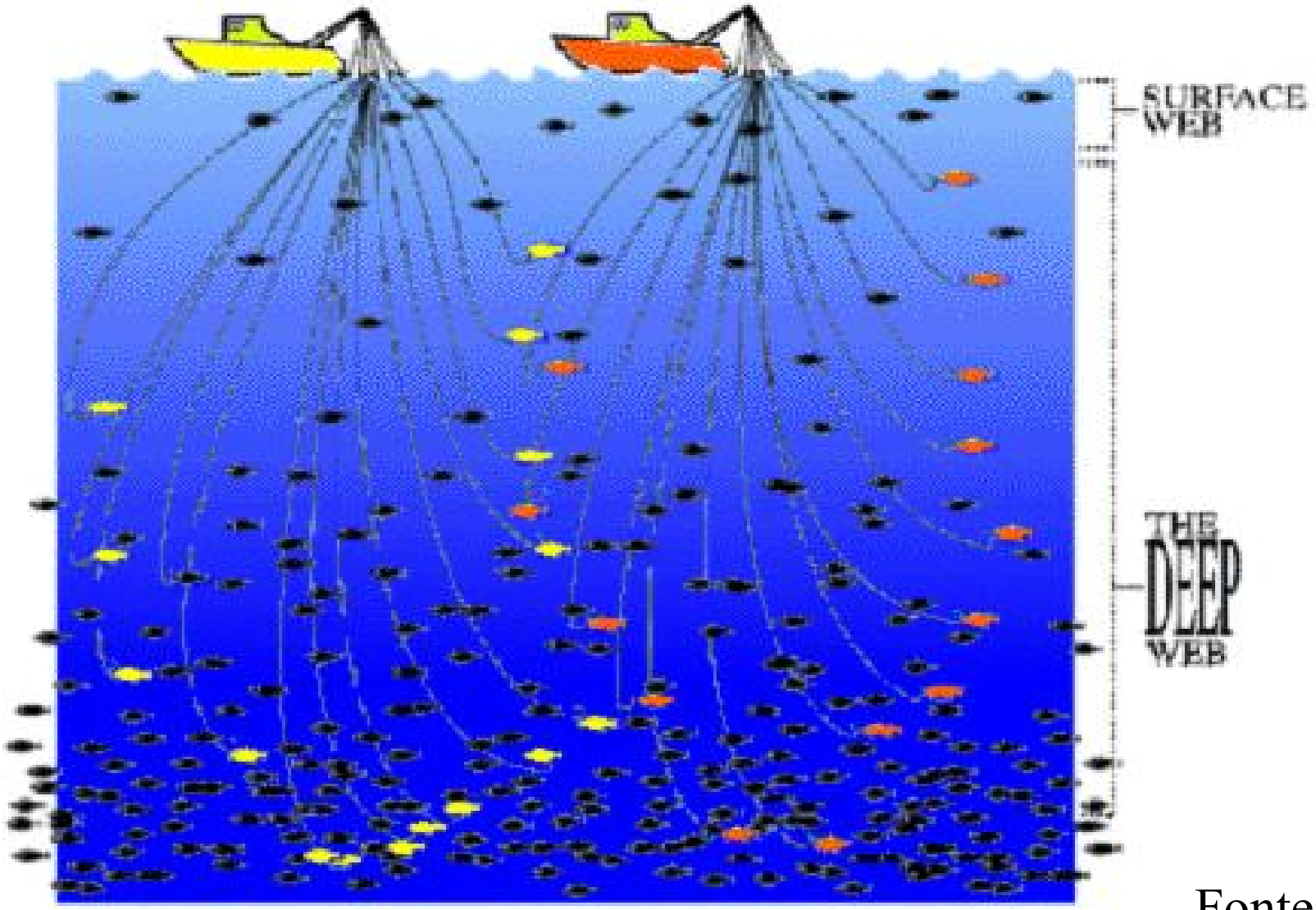
→SOLO UNA PICCOLA PARTE DELLE INFO IN INTERNET E’ DIRETTAMENTE RAGGIUNGIBILE DAL WEB

Quando si naviga in Internet o si usa un motore, si raggiunge generalmente solo una piccola parte delle risorse informative disponibili in Internet: il “surface web”



Fonte: Brightplanet

La parte più consistente delle informazioni è contenuta all'interno del "deep Web"



Fonte: Brightplanet

Il “deep Web”

- Il livello più “interno” dell’informazione reperibile in Internet e/o tramite il World Wide Web:
  - i database e le banche dati accessibili da Internet e/o dal Web
  - i file interni dei server
  - ecc.

I problemi dei motori di ricerca (3/5)

- MODALITÀ DI RICERCA TROPPO SEMPLICI
  - analisi delle ricorrenze non sufficiente
  - insufficiente gestione dei “link”
  - insufficiente gestione della varietà di formati
  - insufficiente gestione della lingua

- ANCHE QUANDO SI USANO ALGORITMI SOFISTICATI:
  - la qualità dei risultati non è garantita
  - l’affidabilità dei risultati non è garantita
  - non si conosce la relazione tra processo di ricerca e risultati!

I problemi dei motori di ricerca (4/5)

- DIFFICOLTÀ DI RICERCA DA PARTE DELL'UTENTE
  - nell'esplicitare/formalizzare il bisogno informativo
  - nella messa a punto della "strategia di ricerca"
  - per l'assenza di procedure "ottimali" o standard

I problemi dei motori di ricerca (5/5)

- AFFIDABILITA' DELLE FONTI E QUALITA' DEI RISULTATI
  - Come essere certi di aver trovato un "buon" risultato?

Se i motori di ricerca non risolvono (del tutto) il problema, cosa si può fare?

- La risposta tecnologica?
  - analisi del linguaggio naturale
  - agenti “intelligenti” di ricerca
  - il “semantic Web”
  - ...
- Al momento non sembra la soluzione

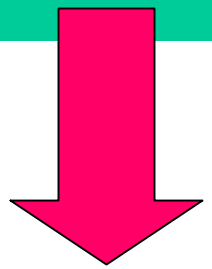
QUINDI?

Strategie (diverse) dei motori di ricerca

- Specializzazione (es. “www.scirus.com”)
- Integrazione - es. google

http://www.google.com/intl/en/options

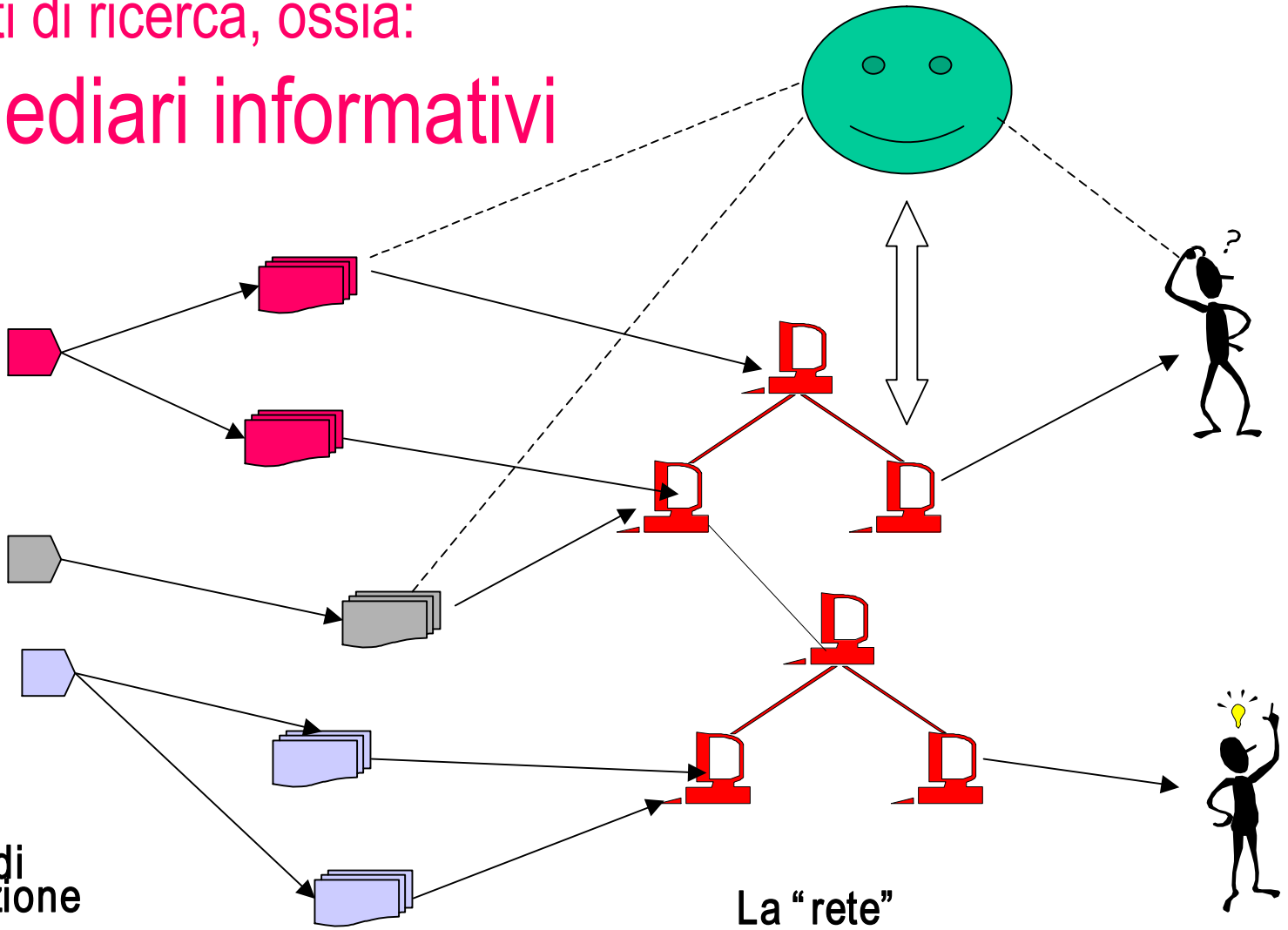
- o ancora sponsorizzazione
- ….



MOTORI DI RICERCA COME “OPERATORI BUSINESS” (INTERMEDIARI INFORMATIVI)

Strumenti di ricerca, ossia:

Intermediari informativi



Le fonti di informazione

La “rete”

Gli utenti

Le risorse informative

Gli intermediari informativi

- Funzione: facilitare l'interazione domanda-offerta di info
  - rendere visibile delle risorse/fonti di informazione
  - facilitare il reperimento da parte degli utenti
- Intermediari come operatori business
- Varietà di situazioni $\Rightarrow$ varietà di operatori
- Catene di intermediari

Quale strumento per quale ricerca?

www.altavista.com

MOTORI DI RICERCA

- ricerca “banale” (ad es. analisi delle ricorrenze)
- come scegliere le parole chiave?
- servizi aggiuntivi (es. traduzioni, cache): quale qualità?
- fonti integrate (mappe, foto, news, libri, …)
- siti sponsor?

www.mamma.com

- Metamotori
  - motori di motori (es.: *mamma*, *metacrawler*)
  - stesse problematiche dei motori, con un'estensione della copertura
  - ricerca più superficiale

www.yahoo.com

DIRECTORY:
- siti preclassificati
- ricerca interna tematica
- sono esaustive?
- come sono selezionati i siti? (sponsor?)

www.istat.it

- Portali informativi
  - Porte di accesso ai depositi di informazione interni (di un'organizzazione, ente, impresa, servizio, ecc.)
  - Ricerca specifica e delimitata
  - Spesso considerata "autorevole" ("fonte ufficiale")
  - Anche "deep Web"
  - Adeguatezza alla ricerca?
  - È davvero affidabile?

www.2night.it

PROBLEMI:
- I PORTALI SONO MOLTO DIVERSI TRA LORO
- BISOGNA CONOSCERLI

- **YELLOW PAGE**
  - Deep Web
  - Ricerca limitata e specifica (elenchi)
  - Ritenuti affidabili (??)
  - Quale copertura? Adeguatezza alla ricerca?
  - Come sono classificate le informazioni?
  - Sono sponsorizzati?
  - Come si effettua la ricerca?

www.paginegialle.it

www.kelkoo.com

- SHOPBOT (SITI DI COMPARAZIONE)
  - ricerche di prodotti/venditori (online);confronti di prezzi/prestazioni (prodotti “consumer”)
  - informazioni altamente specifiche
  - quale classificazione?
  - sponsorizzazione?

www.cab.unipd.it

BANCHE DATI
- deep web
- altamente specifiche
- generalmente (ritenute) affidabili
- bisogna conoscerle
- bisogna saper cercare

www.profinder.it

- SERVIZI CON OPERATORE
  - assistenza alla ricerca
  - per chi non ha tempo (o non è competente)
  - validità della risposta?
  - Affidabilità?
  - Costo?

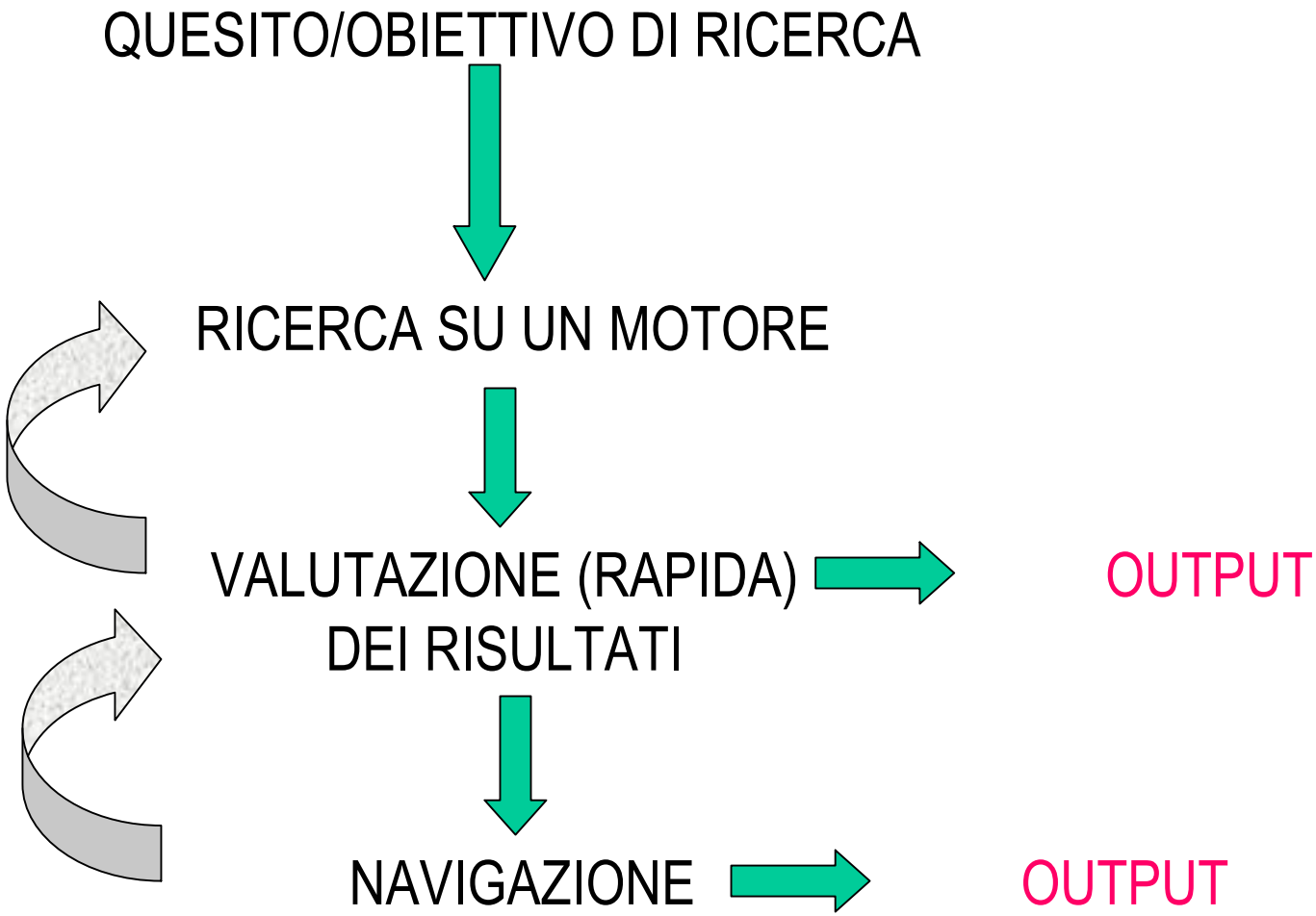
- La navigazione diretta
  - necessario conoscere in anticipo l'indirizzo
    - pre-selezione da altre fonti (non in rete)
    - metodo "snowball"
    - come reperire all'interno del sito?
      - » NAVIGAZIONE LIBERA
      - » MOTORI DI RICERCA INTERNI
  - Siti "autorevoli"? Ufficiali?

Come impostare una ricerca su Internet?

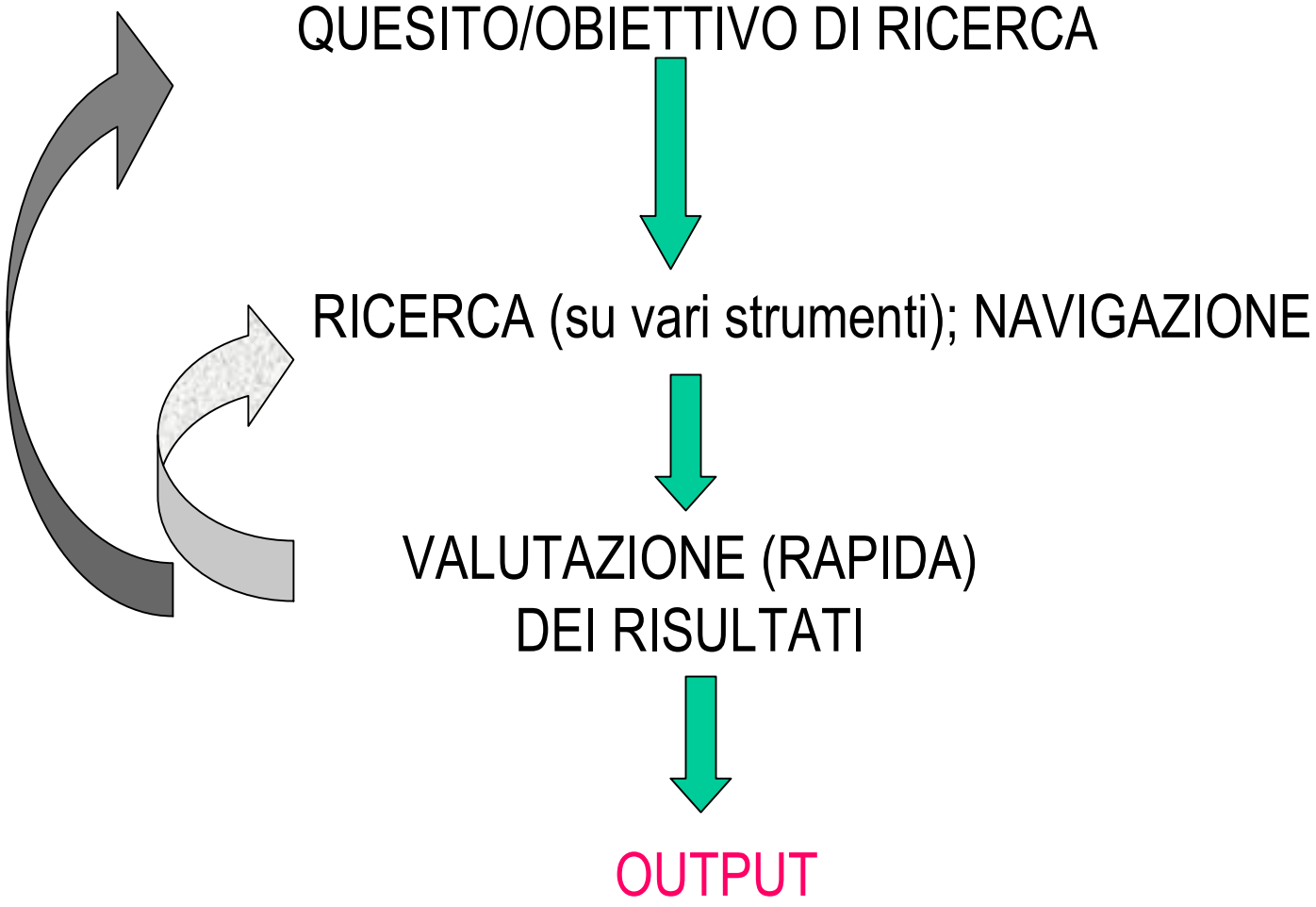
La capacità dell'utente

- Information literacy: imparare a cercare su Internet
- Oggi:
  - utenti sempre più vari
  - non competenti del mezzo Internet
  - approccio di ricerca "intuitivo"
- L'importanza della "prior knowledge"

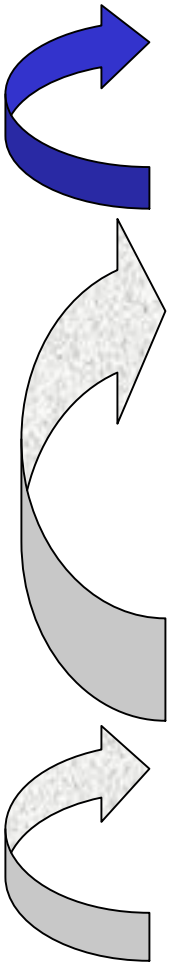
Approcci alla ricerca: navigare



Approcci alla ricerca: il “berry picking”



Approcci alla ricerca: cercare analiticamente



QUESITO/OBIETTIVO DI RICERCA

IMPOSTAZIONE:

- identificazione del problema e del contesto cognitivo
- identificazione e scelta di concetti, nozioni, parole chiave
- scelta degli strumenti di ricerca più opportuni

RICERCA (su vari strumenti)

COMPARAZIONE E VALUTAZIONE DEI RISULTATI → OUTPUT

Impostare una ricerca: elementi

1. Il quesito, l'oggetto, l'obiettivo della ricerca ("cosa" e "perché")
2. Il tempo e la capacità di chi cerca
3. La qualità delle risposte
   - Adeguatezza agli obiettivi/domande di ricerca
     - Completezza – esaustività
     - oppure: focalizzazione
   - Affidabilità/autorevolezza della fonte; modalità di verifica
   - Grado di aggiornamento

LA PROSPETTIVA DELLE FONTI DI INFORMAZIONE: COME RENDERE VISIBILI LE PROPRIE PAGINE?

- Conoscendo il funzionamento dei motori di ricerca
  - Ad es.: includere le parole che si vuole siano indicizzate; ripetere più volte le parole (si deve ragionare sul modo di cercare degli utenti nel motore di ricerca)
  - Evitando di mettere le informazioni chiave nelle sotto-sotto-sotto pagine
- “Pagando” per una migliore collocazione
- Ricorrendo ad altri servizi (es. banner o link in “siti di traffico”)