UNIVERSITÀ DEGLI STUDI DI FERRARA
- EX LABORE FRUCTUS -

TECNOLOGIE INFORMATICHE E MULTIMEDIALI
AA 2016-2017

CORSO DI LAUREA: «*SCIENZE E TECNOLOGIE DELLA COMUNICAZIONE*»

Prof. Giorgio Poletti
giorgio.poletti@unife.it

«GOOGLE È UN GLOBALE TEST DI RORSCHACH. VEDIAMO IN ESSO CIÒ CHE VOGLIAMO VEDERE.»

John Battelle

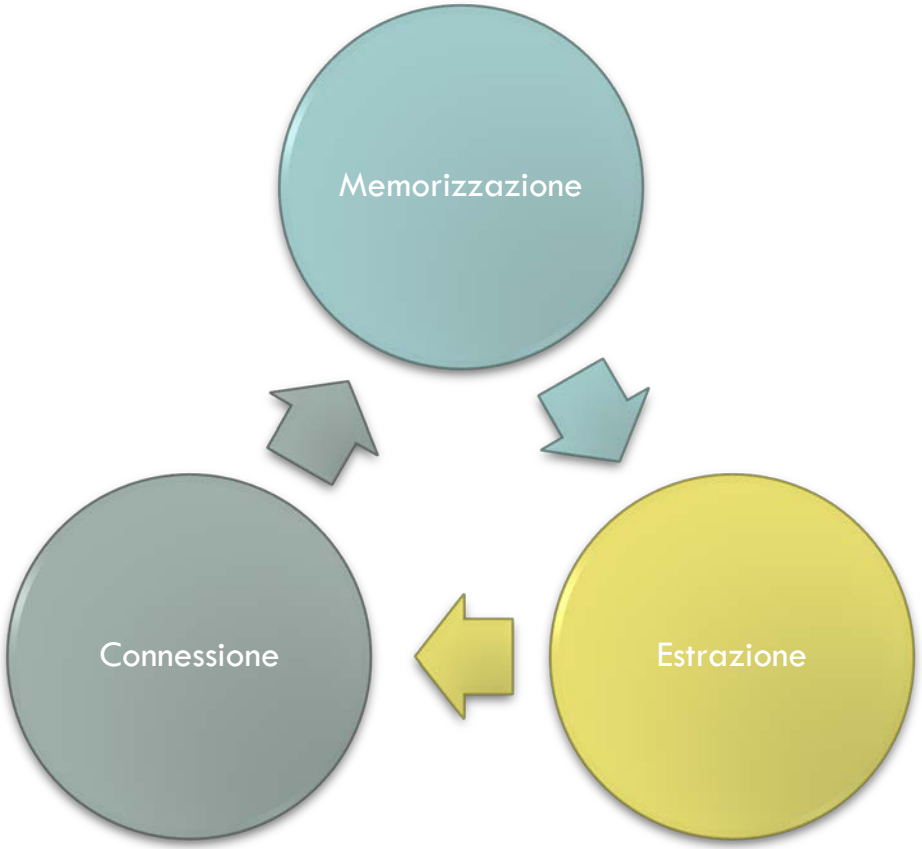
Scrittore e giornalista 1965-

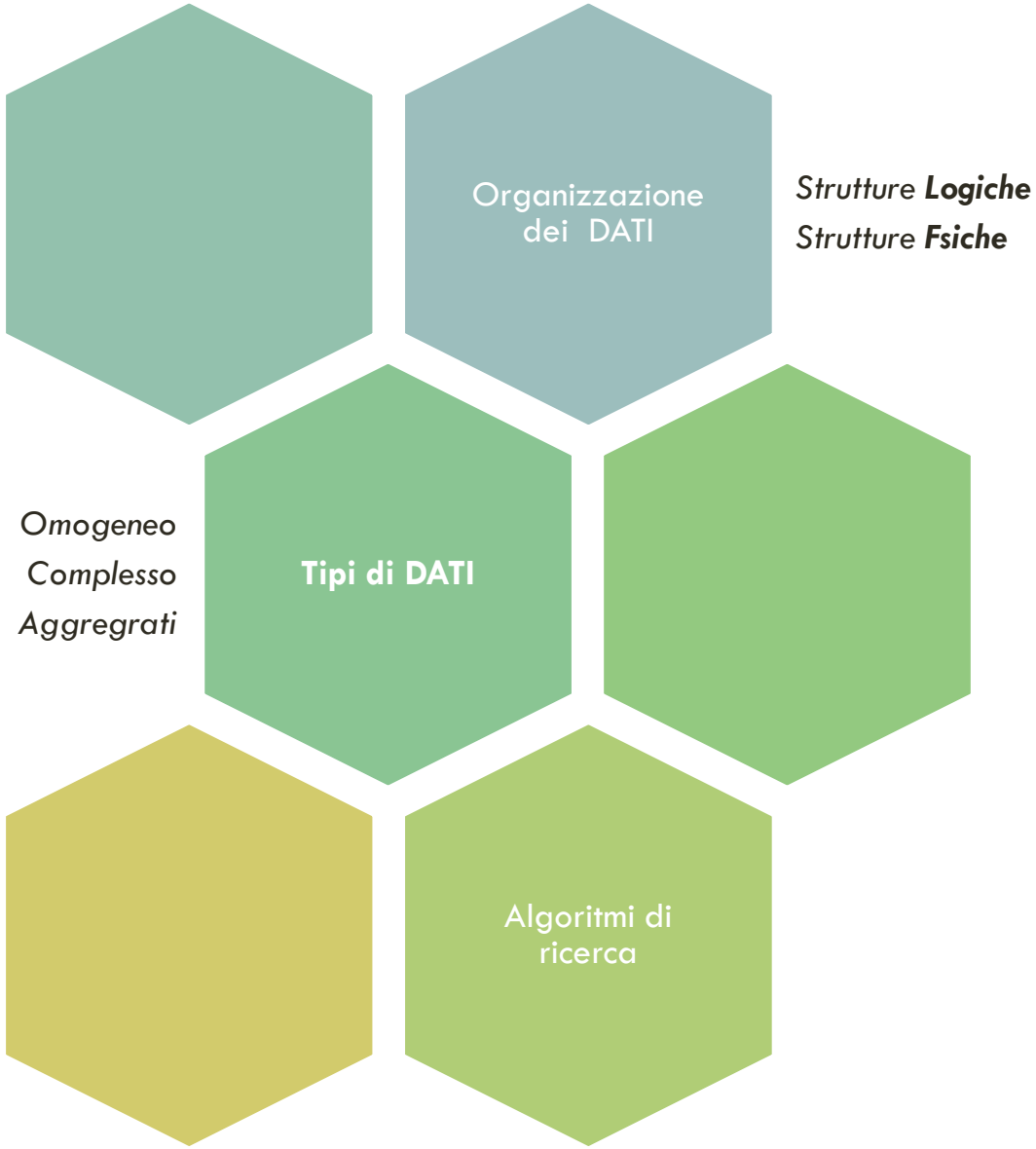
DATA BASE E MOTORI DI RICERCA

Memorizzare i dati dall'archivio al DataBase. Cercare n Rete

INFORMATION MANAGEMENT

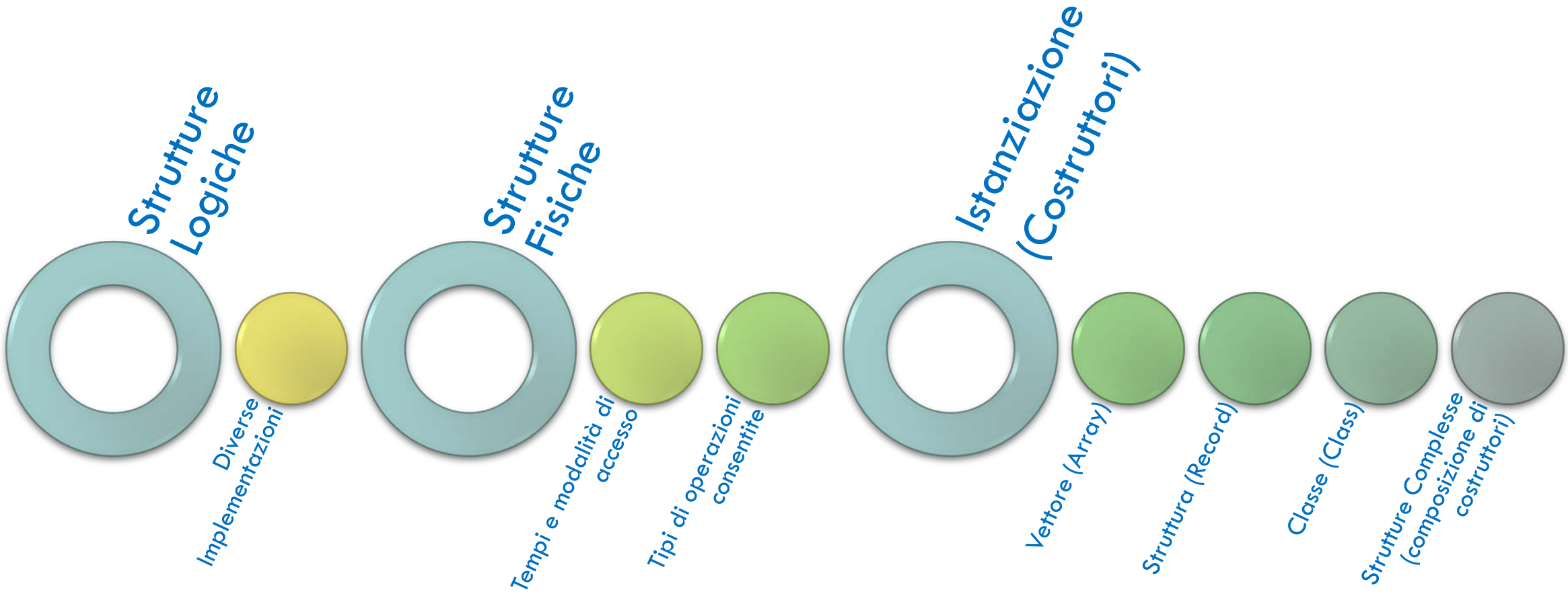
Astrazione per progettare





INFORMATION MANAGEMENT

Strutture Dati



INFORMATION MANAGEMENT

Strutture Dati

VETTORE (ARRAY)

- Dati omogenei
- Numero finito di elementi
- Lunghezze diverse oggetti diversi
- Costo computazionale
  - COSTANTE in accesso
  - inserimento e cancellazione ONEROSO

STRUTTURA (RECORD)

- Dati omogenei o eterogenei
- Gli elementi sono detti CAMPI

CLASSE (CLASS)

- Record con associati metodi (azioni)

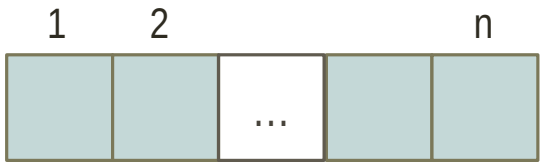
STRUTTURE COMPLESSE (COMPOSIZIONE DI COSTRUTTORI)

- Costruiti come combinazioni dei costruttori Vettore, Struttura e Classe e sono STRUTTURE STATICHE

INFORMATION MANAGEMENT

Strutture Dati

Strutture Statiche

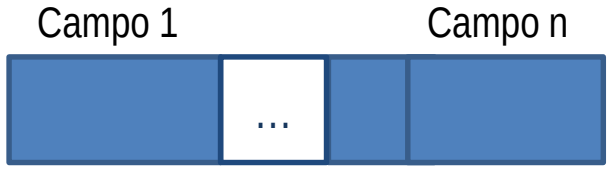


Strutture Dinamiche

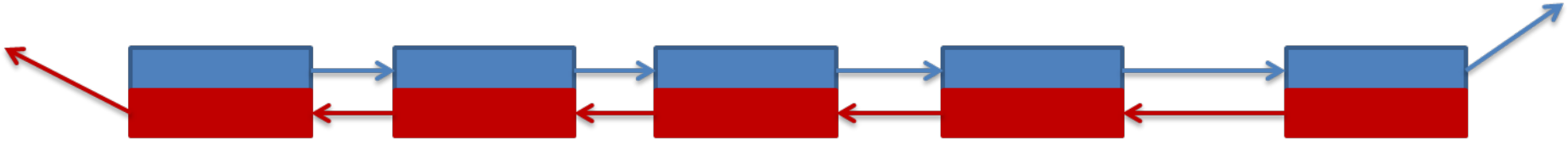
PUNTATORI

ALLOCAZIONE DINAMICA della Memoria

- **LISTA (Catena)**



- **LISTA BIDIREZIONALE (Doppia Catena)**

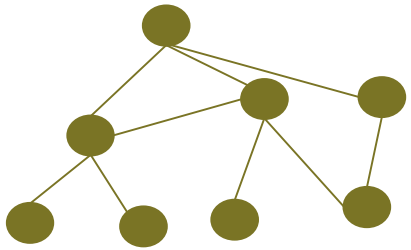


INFORMATION MANAGEMENT

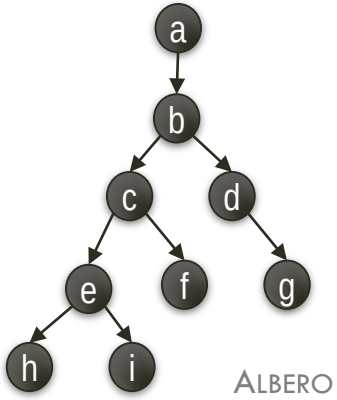
Strutture Dati

Strutture Dinamiche

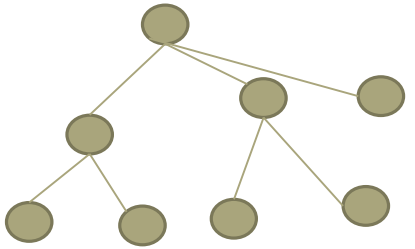
- ✓ ALBERO Ordinato
- ✓ ALBERO Binario Ordinato
- ✓ GRAFO



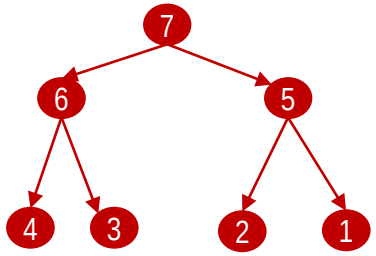
GRAFO



ALBERO ORDINATO



ALBERO



ALBERO BINARIO ORDINATO

INFORMATION MANAGEMENT

Strutture Dati

Strutture Dinamiche

- ✓ TABELLA HASH (hash, pezzettino, table)

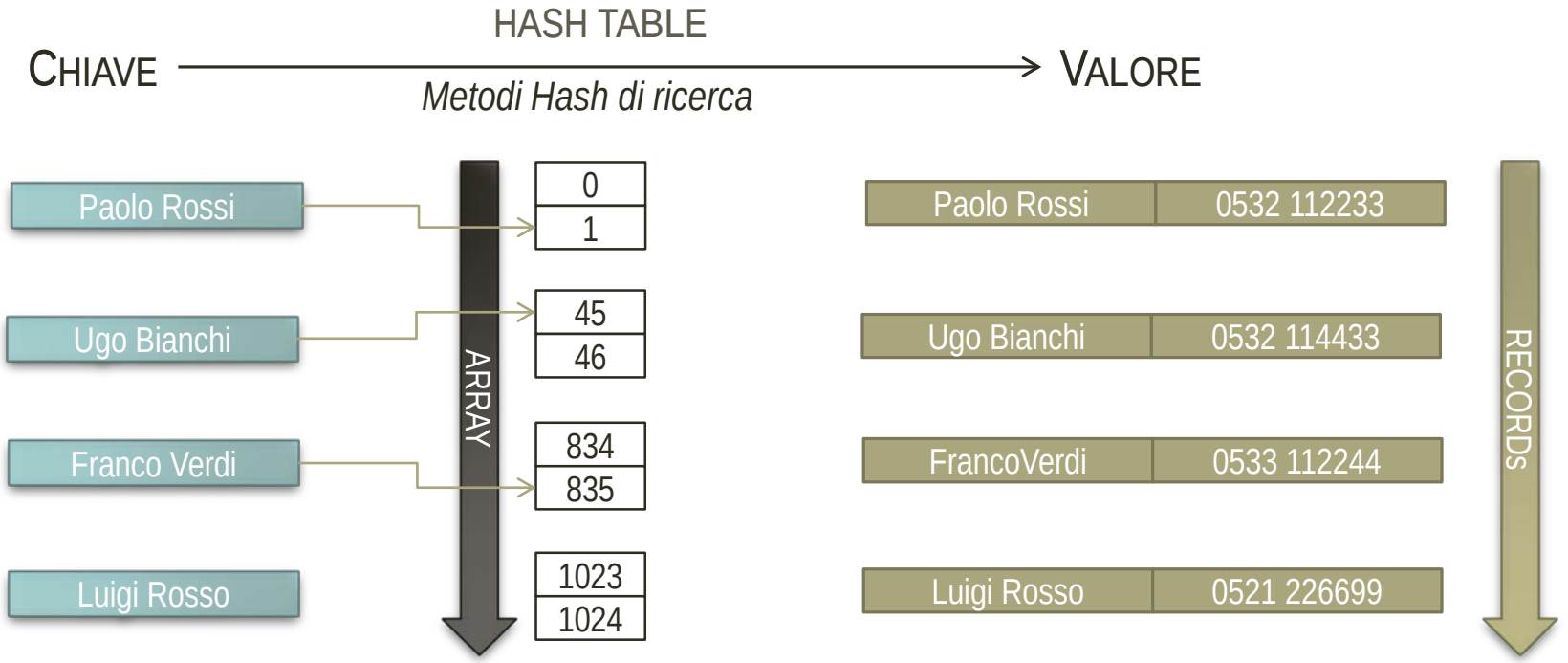
In senso figurato rappresenta un composto eterogeneo dalla forma incerta, e in informatica la funzione ***hash*** *è una funzione che mappa una stringa di lunghezza arbitraria in una stringa di lunghezza predefinita. Struttura dati utile per cercare velocemente un elemento all'interno di un insieme sulla base di una chiave*

INFORMATION MANAGEMENT

Strutture Dati

Strutture Dinamiche

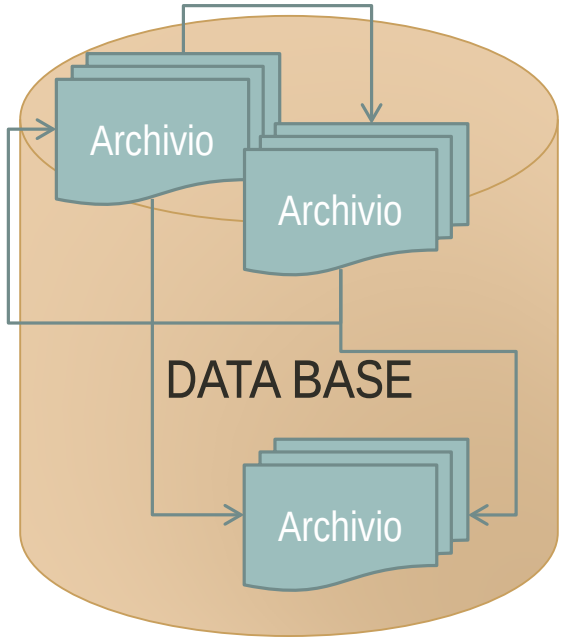
- ✓ TABELLA HASH - Esempio



INFORMATION MANAGEMENT

Data Base

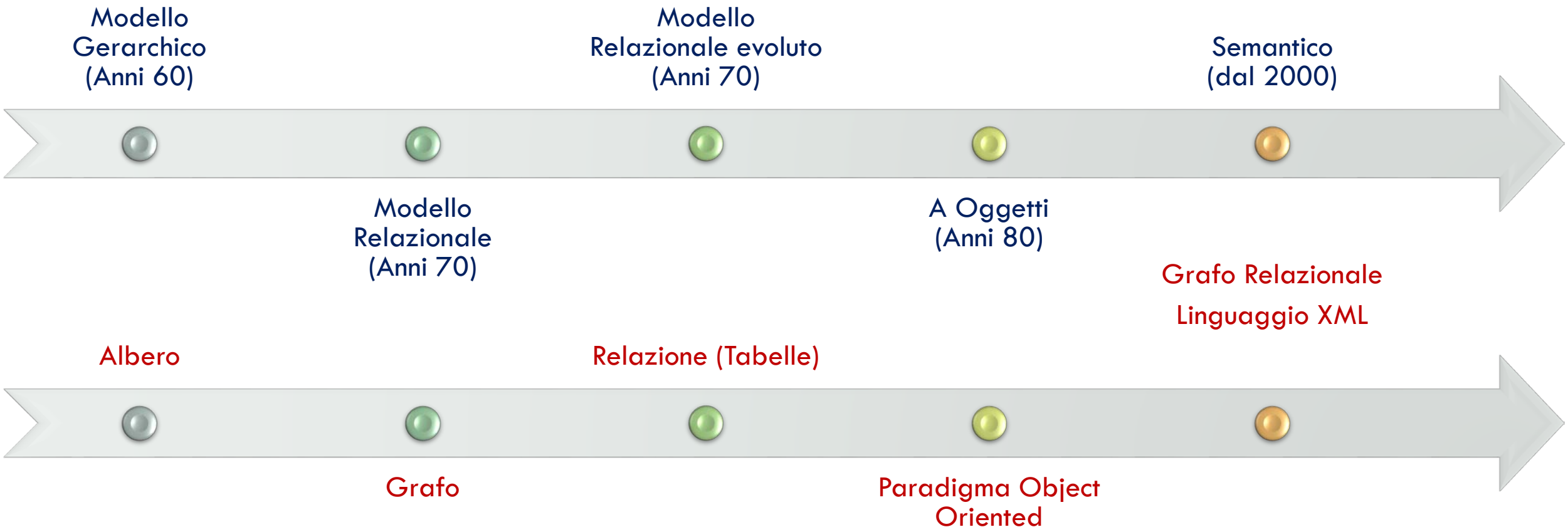
- ✓ INSIEME DI ARCHIVI
- ✓ MODELLO LOGICO DI CONNESSIONE



Dati Correlati → Informazioni Reperibili

INFORMATION MANAGEMENT

Data Base – Time Line (Modello Logico/Modello di Rappresentazione)



INFORMATION MANAGEMENT

Data Base – Time Line (Modello Logico/Modello di Rappresentazione)

GERARCHICO/ALBERO

- Gerarchia tra entità
- Una Radice e uno o più sottoalberi dipendenti
- Struttura 1:N (padre:figli)

- Albero come schema logico
- Scarsa indipendenza dello schema logico
- Gli schemi esterni devono esplicitare tutto il percorso

RETICOLARI/GRAFO

- Record e Puntatori
- Struttura M:N

- Aumento della quantità di dati e Ridondanza
- Aumento della possibilità di errore

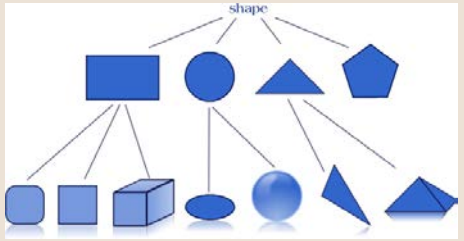
RELAZIONALI/TABELLE

- Tutti i dati sono relazioni e manipolati con l'algebra relazionale
- Struttura M:N

- Problemi nella modellazione di relazioni complesse

OGGETTI

- Paradigma Object Oriented: memorizzazione dei dati come «oggetti riutilizzabili».
- Oggetto come insieme di diversi tipi di dati (testo, audio, video) e metodi (operazioni eseguibili sui dati).



INFORMATION MANAGEMENT

DBMS (Data Base Management System) - CARATTERISTICHE

Sistema di controllo che permetta la gestione di grandi masse di dati attraverso uno strumento centralizzato.

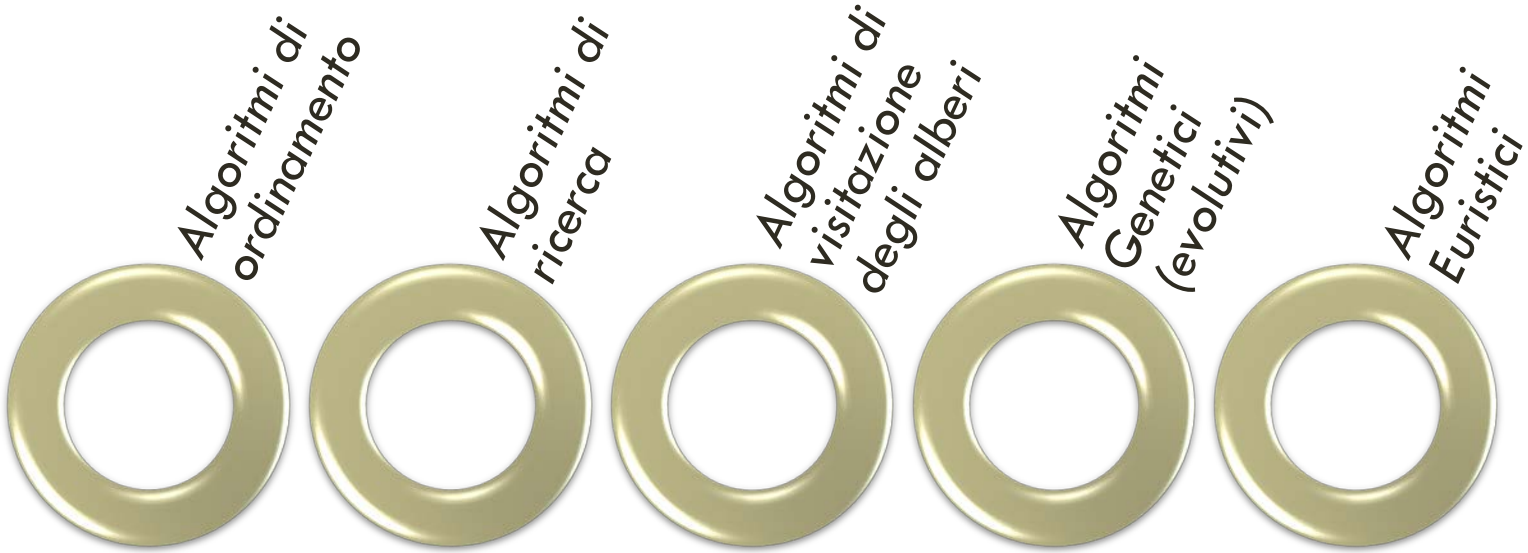
- Controllo ridondanza e consistenza
- Accessi simultanei ai dati
- Integrità dei testi
- Sicurezza delle Informazioni
- Facile accesso ai dati per non esperti

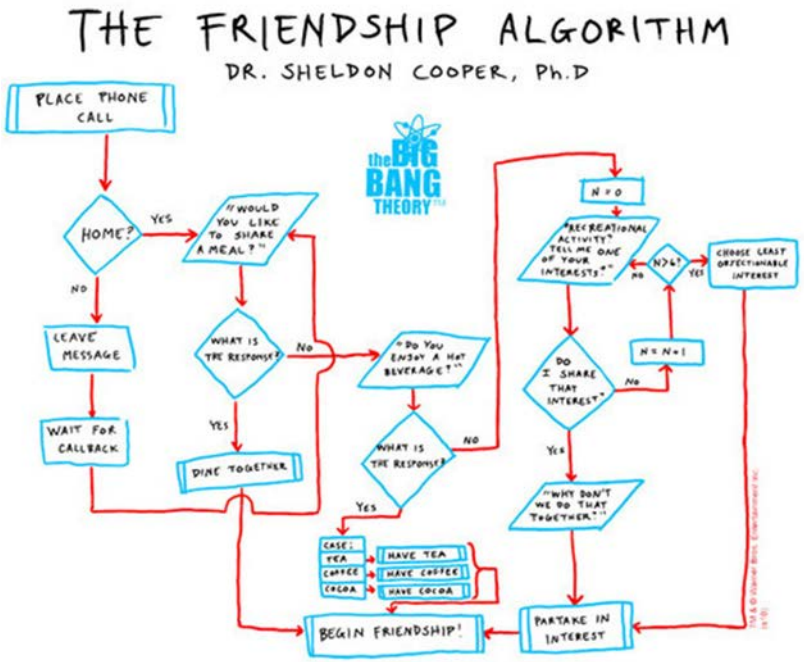
CONSISTENZA: *sinonimo di coerenza, conformità, non contraddittorietà, affidabilità dei dati.*

Un dato o una informazione si dicono INTEGRI se sono rimasti inalterati nel loro contenuto, durante la loro trasmissione oppure la loro memorizzazione.

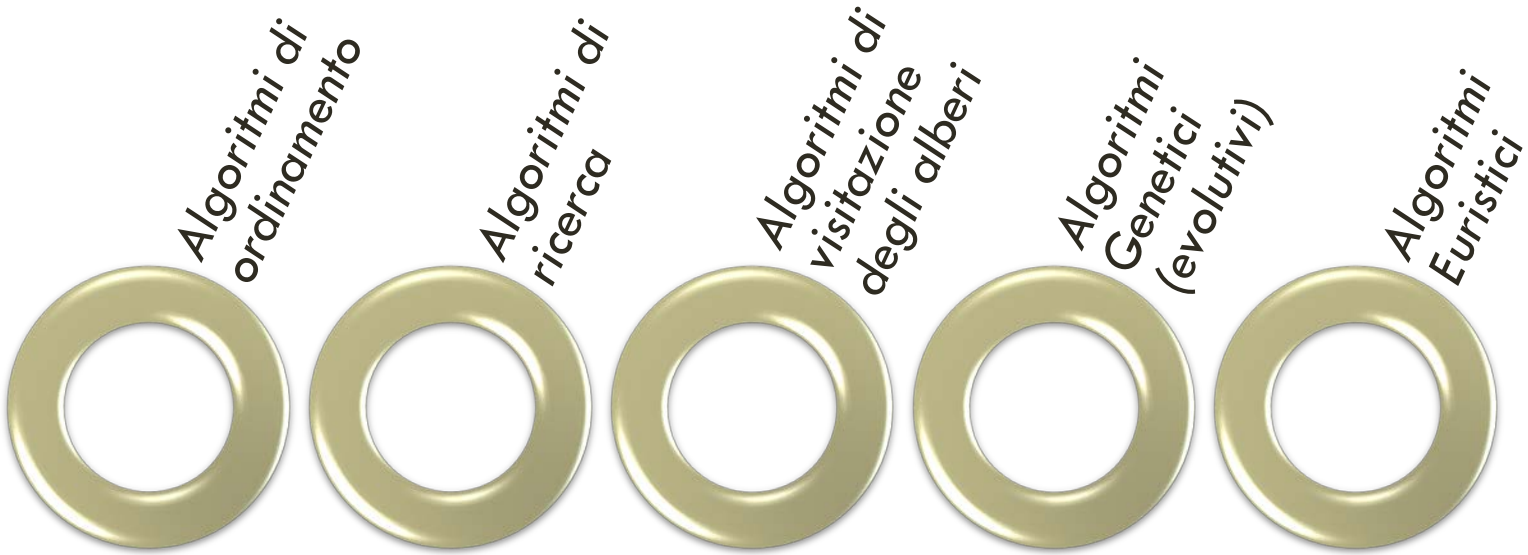
ALGORITMI DI RICERCA DELLE INFORMAZIONI

Molteplici e rendono impossibile una catalogazione precisa e definitiva





ALGORITMI DI RICERCA DELLE INFORMAZIONI



OTTIMALITÀ

Trova sempre la soluzione migliore: quella meno profonda (p minore), **è *OTTIMALE***

COMPLETEZZA

Trova sempre la soluzione se c'è, **è *COMPLETO***

ALGORITMI DI RICERCA DELLE INFORMAZIONI

Algoritmi Genetici

CROMOSOMA: *una delle soluzioni ad un problema considerato e generalmente è codificata con un vettore di bit o di caratteri*
POPOLAZIONE: *insieme di soluzioni relative al problema considerato*
GENE: *parte di un cromosoma e generalmente consiste in una o più parti del vettore di bit o caratteri che codificano il cromosoma*
FITNESS: *grado di valutazione associato ad una soluzione e la valutazione avviene in base ad una funzione appositamente progettata detta funzione di fitness*
CROSSOVER: *generazione di una nuova soluzione mescolando delle soluzioni esistenti*
MUTAZIONE: *alterazione casuale di una soluzione*

ESEMPIO DI ALGORITMO GENETICO (PROVVEDE A FARE EVOLVERE DELLE SOLUZIONI)

- **GENERAZIONE** *casuale della prima popolazione di soluzioni (cromosomi).*
- **APPLICAZIONE** *della funzione di fitness alle soluzioni (cromosomi) appartenenti all'attuale popolazione.*
- **SELEZIONE** *delle soluzioni considerate migliori in base al risultato della funzione di fitness e della logica di selezione scelta*
- **PROCEDIMENTO** *di crossover per generare delle soluzioni ibride a partire dalle soluzioni scelte al punto*
- **CREAZIONE** *di una nuova popolazione a partire dalle soluzioni identificate al punto*
- **RIESECUZIONE** *della procedura a partire dal punto 2 ed utilizzando la nuova popolazione creata al punto*
- **L'ITERAZIONE** *dei passi presentati permette l'evoluzione verso una soluzione ottimizzata del problema considerato*

ALGORITMI DI RICERCA DELLE INFORMAZIONI

Algoritmi Euristici

Usa INFORMAZIONI CARATTERISTICHE del problema per decidere quali stati espendere

All'insieme delle configurazioni (spazio degli stati) associa una FUNZIONE EURISTICA che valuta gli stati e gli assegna un valore (valorizza ogni nodo-stato)

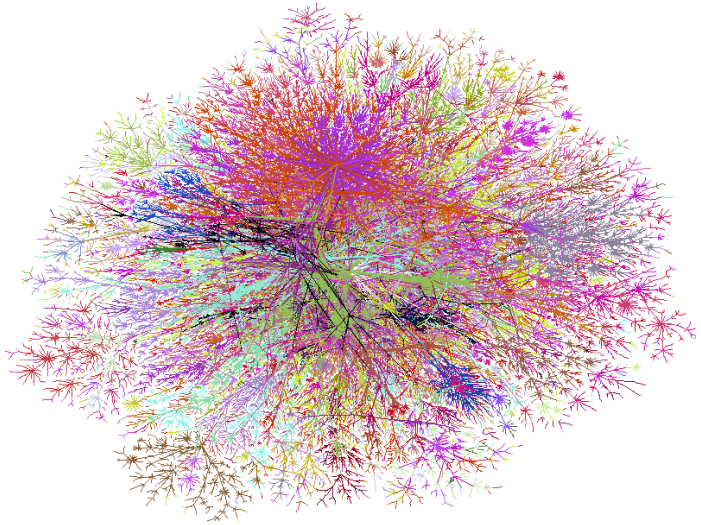
Applica una STRATEGIA EURISTICA: sceglie il nodo più «interessante», con «migliori prospettive» e valutato più vicino al noto target (stato finale)

ALGORITMI DI RICERCA DELLE INFORMAZIONI

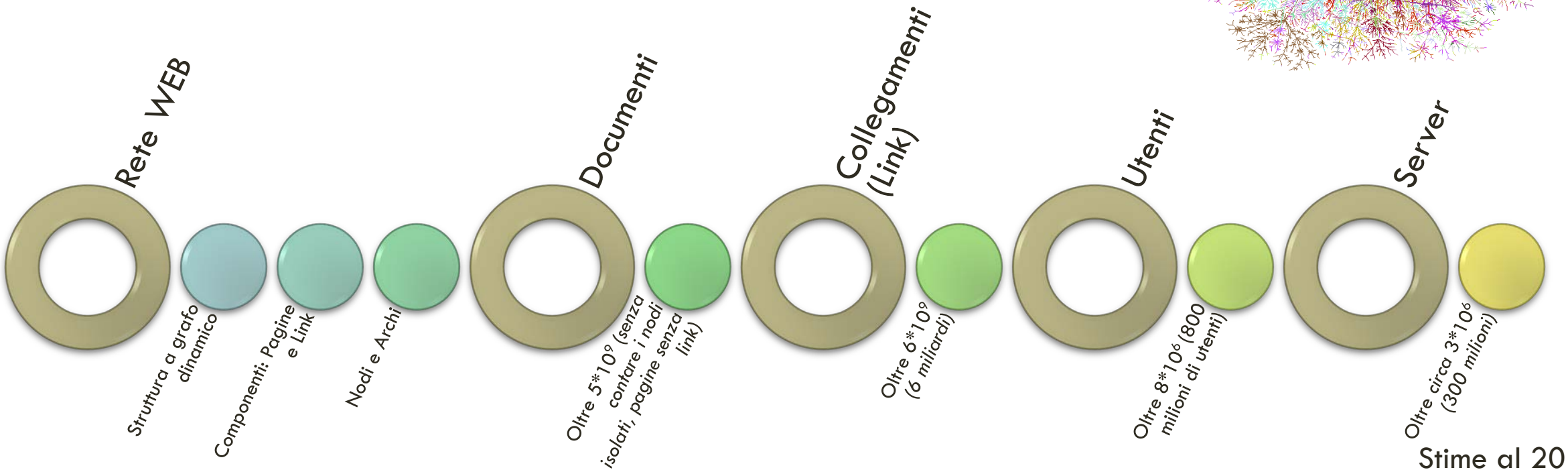
Esempio di costo della ricerca

	COSTO DELLA RICERCA		
p	ID (Iterative Deepening)	A*(h_1)	A*(h_2)
2	10	6	6
4	112	13	12
6	680	20	18
8	6.384	39	25
10	47.127	93	39
12	364.404	227	73
14	3.473.941	539	113

ALGORITMI DI RICERCA SUL WEB



Il PageRank

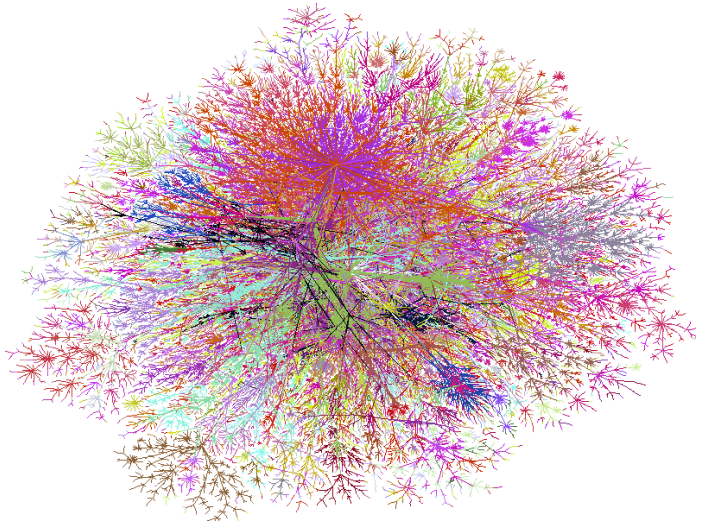


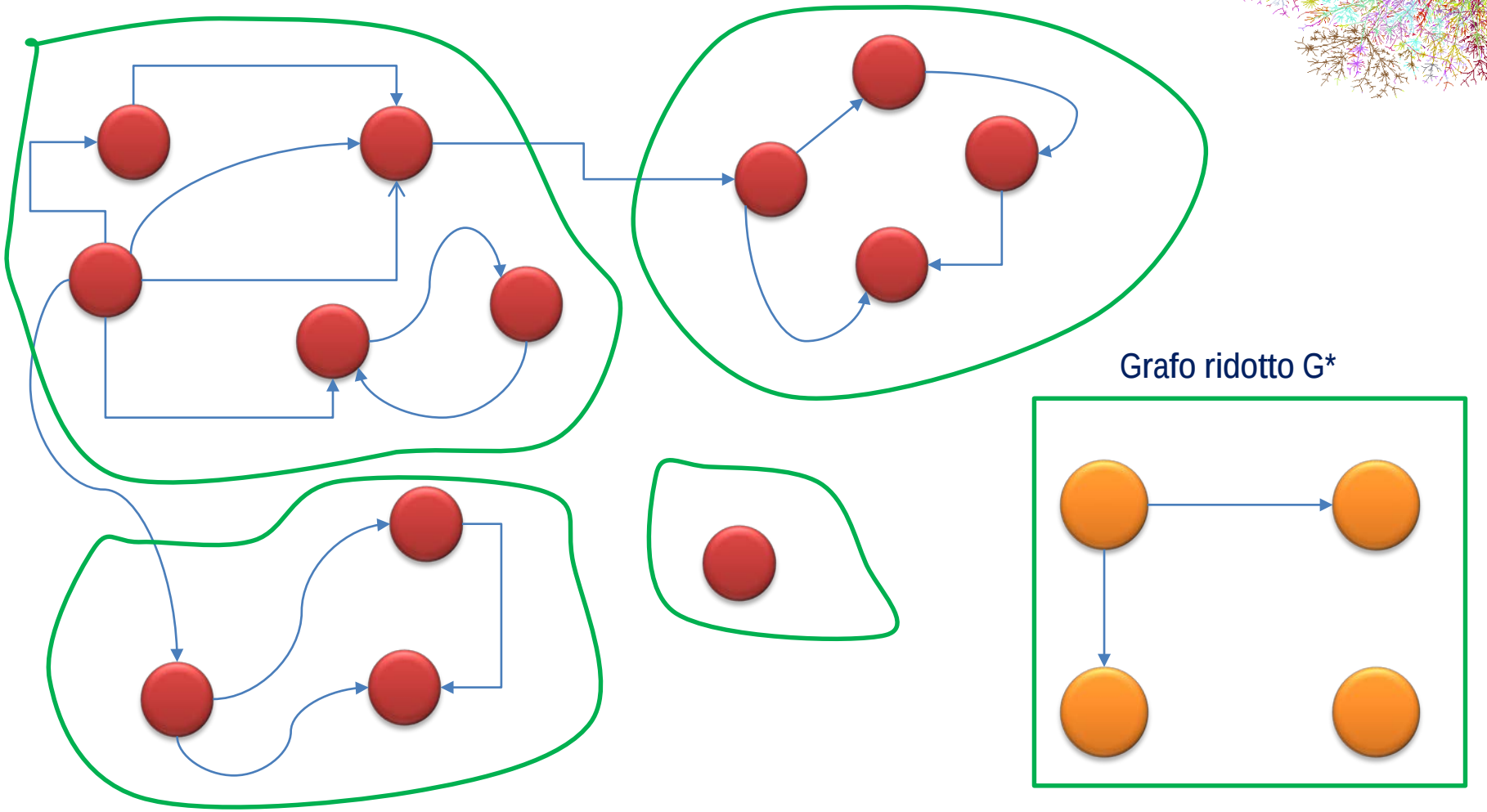
Stime al 2014

DEFINIZIONE: GRAFO RIDOTTO G*, *i nodi sono le classi e due classi C_1 e C_2 sono connesse se esiste un nodo in C1 collegato a un nodo in C2, esiste un arco da C_1 a C_2*

ALGORITMI DI RICERCA SUL WEB

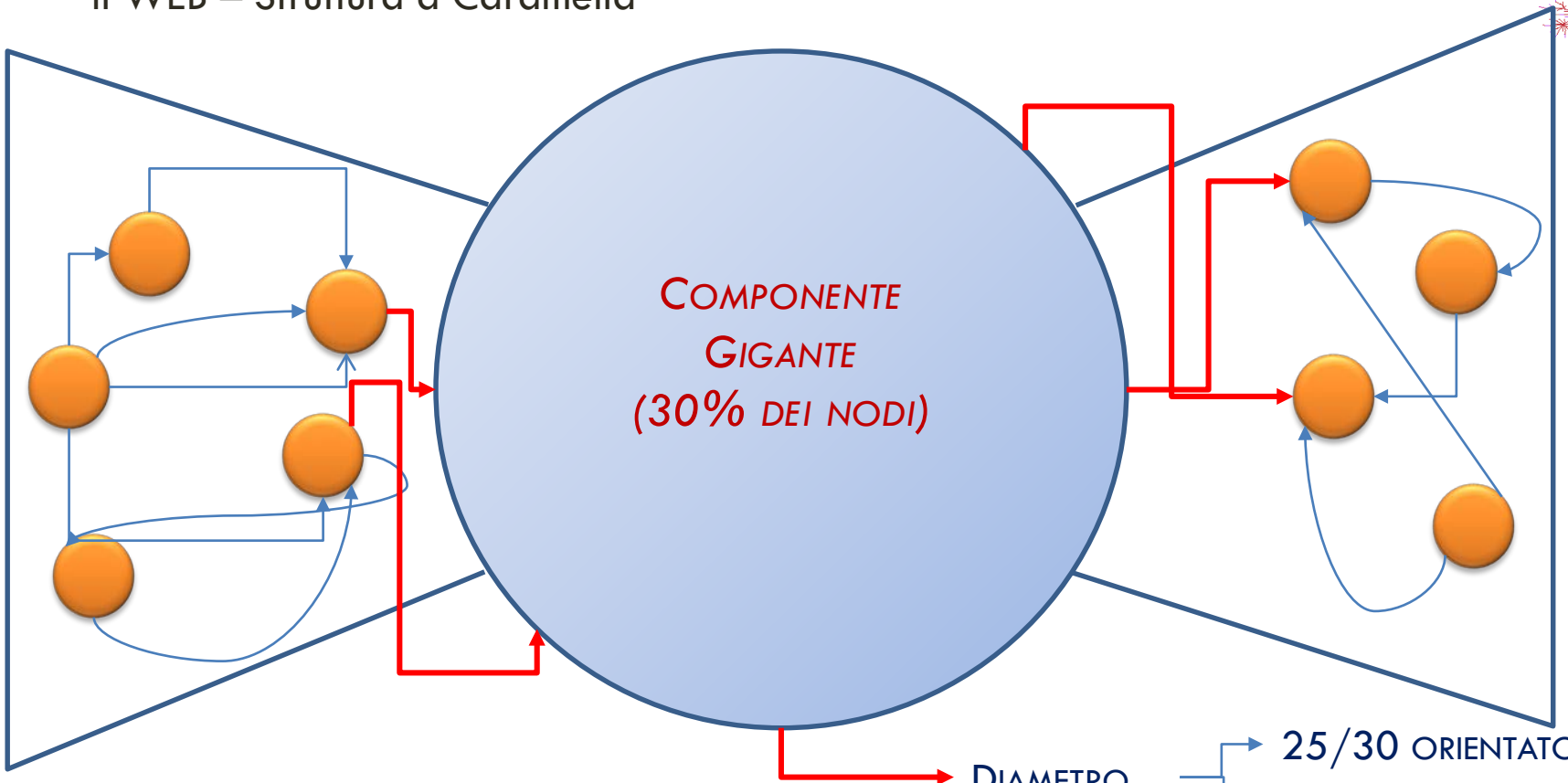
Grafo Ridotto - Esempio

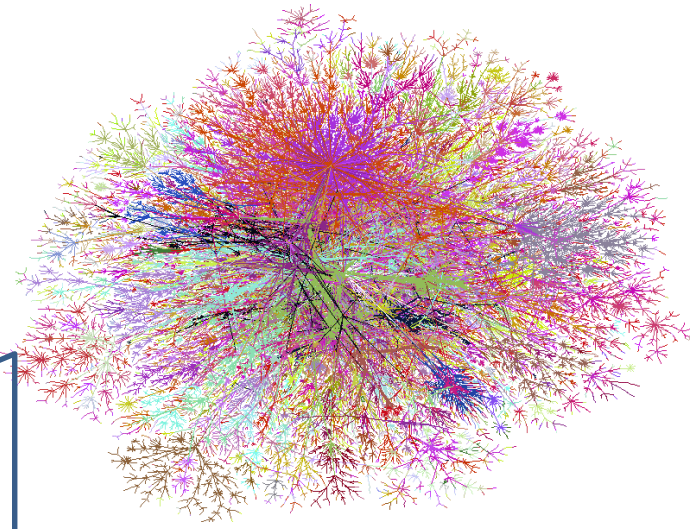




ALGORITMI DI RICERCA SUL WEB

Il WEB – Struttura a Caramella



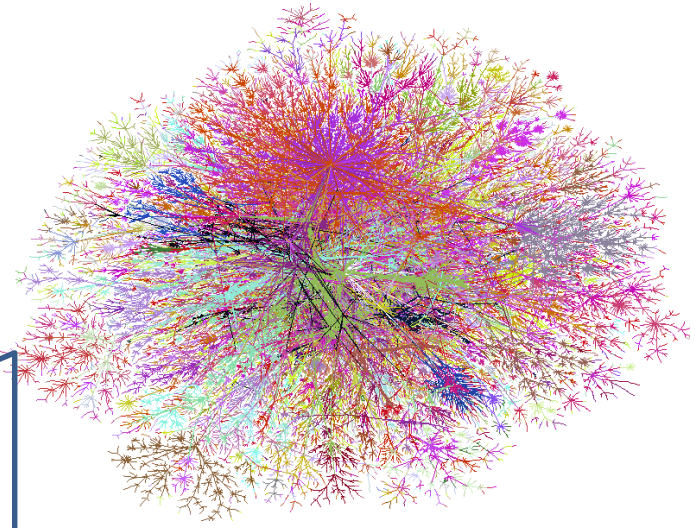


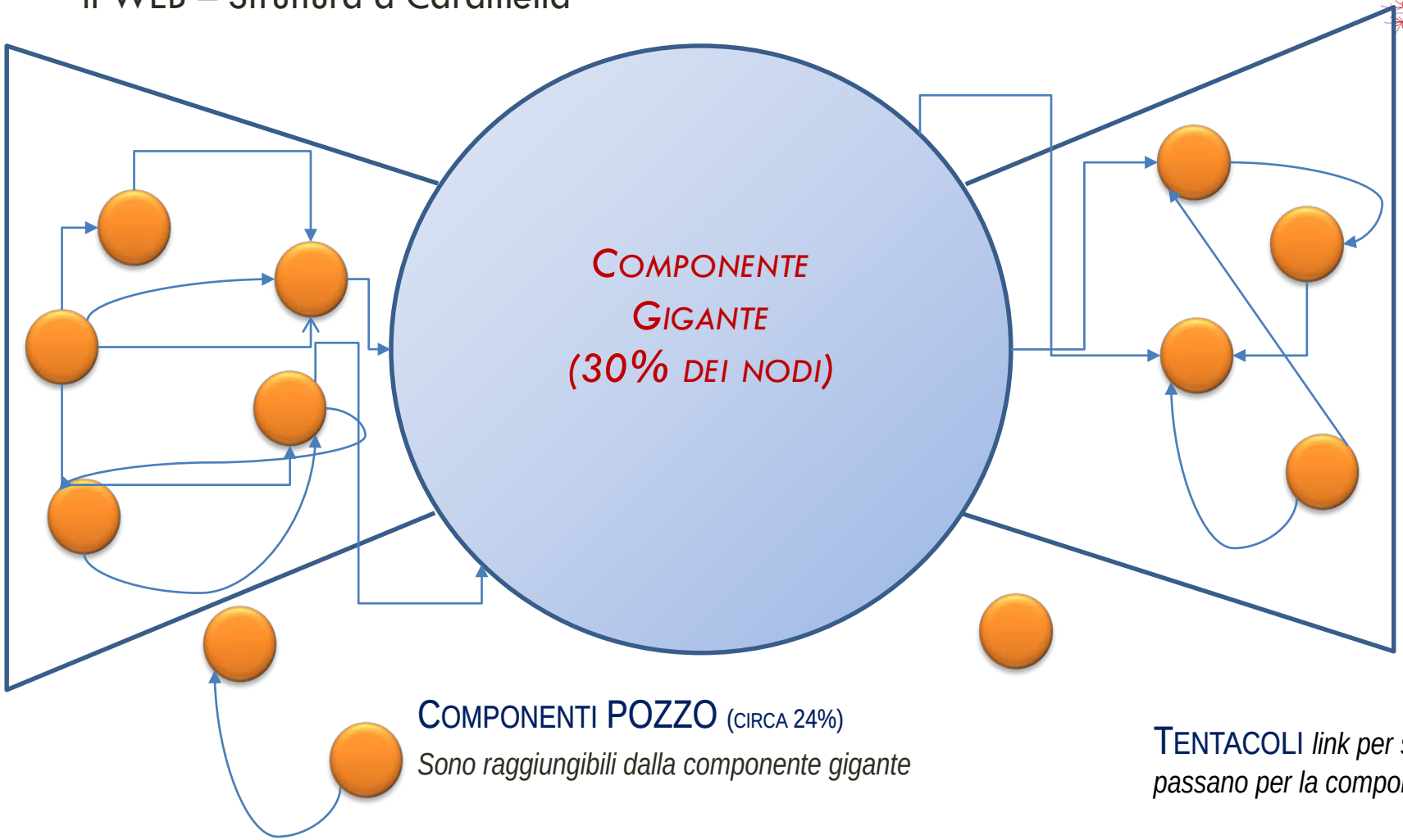
Teoria dei piccoli mondi
(***SMALL WORLD THEORY***)

La teoria dei **piccoli mondi** sostiene che tutte le reti complesse presenti in natura sono tali che due qualunque nodi possono essere collegati da un percorso costituito da un numero relativamente piccolo di collegamenti. (6 gradi di separazione)

ALGORITMI DI RICERCA SUL WEB

Il WEB – Struttura a Caramella





COMPONENTI SORGENTE (CIRCA 24%)

Puntano in maniera diretta o indiretta la componente gigante

PAGINE «REIETTE»

Non sono raggiungibili dalla componente gigante

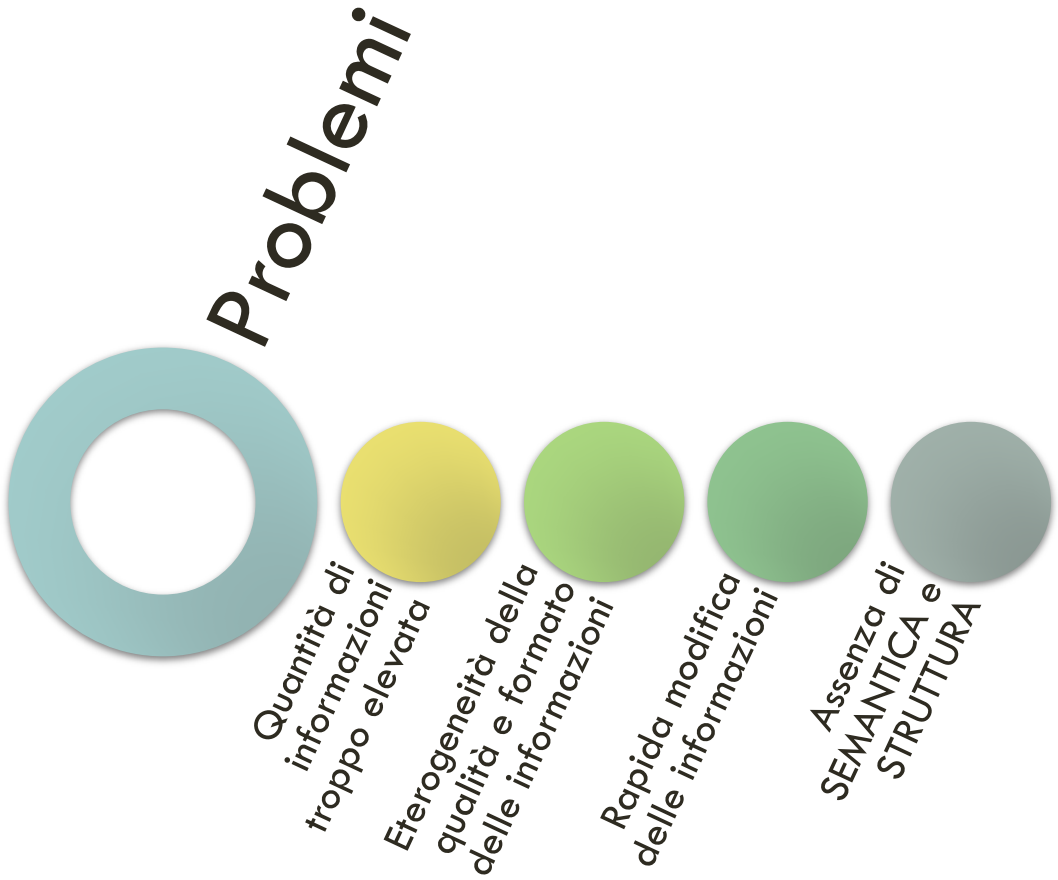
COMPONENTI POZZO (CIRCA 24%)

Sono raggiungibili dalla componente gigante

TENTACOLI *link per sorgenti e pozzi che non passano per la componente gigante*

ALGORITMI DI RICERCA SUL WEB

Motori di Ricerca (Oltre 80% degli utenti li usa)



MOTORI DI RICERCA

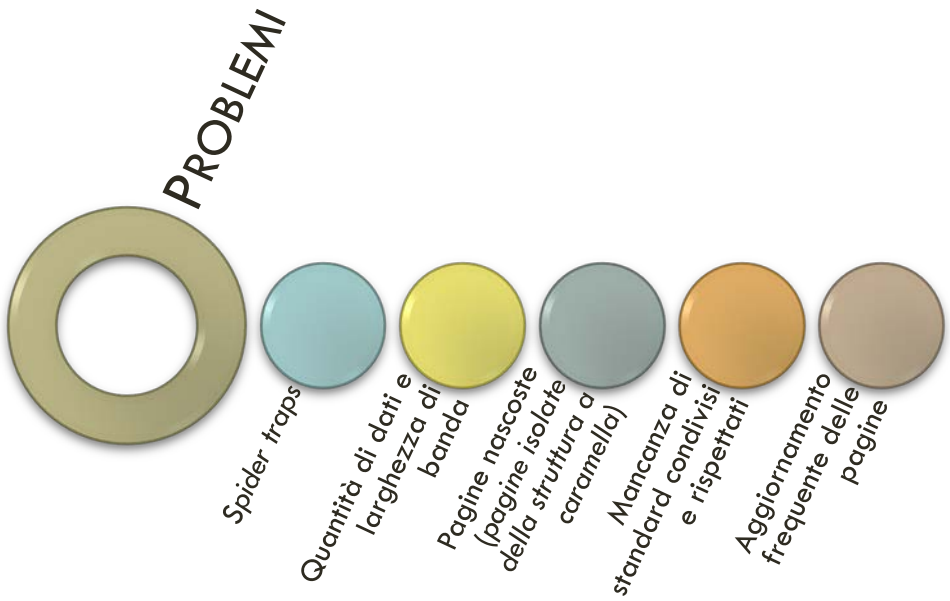
Azioni

MOTORI DI RICERCA

Raccolta dati

Raccolta del contenuto delle pagine Web (informazioni di tipo testuale ma anche immagini (google immagini e documentazione), robot.txt

Si usa uno spider o crawler o robot



SPIDER	MOTORE DI RICERCA
googlebot	Google
fast	Fast - Alltheweb
scooter	Altavista
mercator	Altavista
Ask Jeeves	Ask Jeeves
teoma_agent	Teoma
ia_archiver Yahoo!	Alexa - Internet Archive
Slurp	Yahoo
Romilda	Facebook

MOTORI DI RICERCA

Elaborazione e catalogazione dei dati raccolti

Elaborazione e catalogazione dei dati raccolti

Rilevazione delle ridondanze (presenza di MIRRORING)

Rilevazione di presenza di SPAMMING

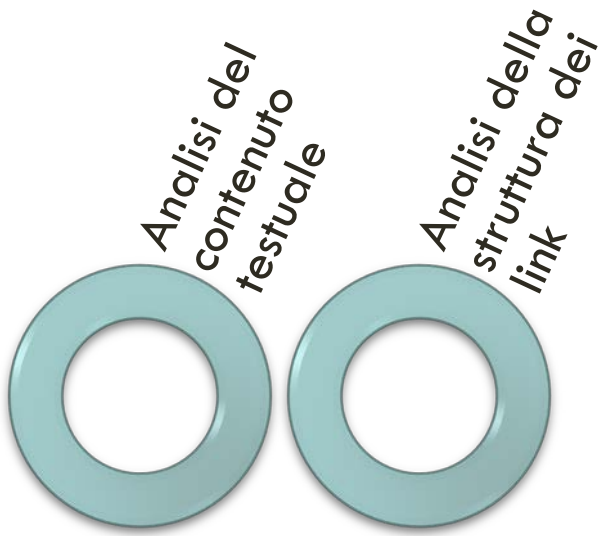
Reperimento e analisi delle informazioni per il calcolo del **RANKING**

ALGORITMO DI PAGERANK

RANKING: Dato un insieme di pagine **P** e una query **Q** il ranking è definito da una funzione:

$R_q: P \rightarrow \mathbf{R}$ *(insieme dei numeri reali)*

*che associa ad ogni pagina un numero reale che indica la «**RILEVANZA**» di quella pagina nel contesto di quella query.*

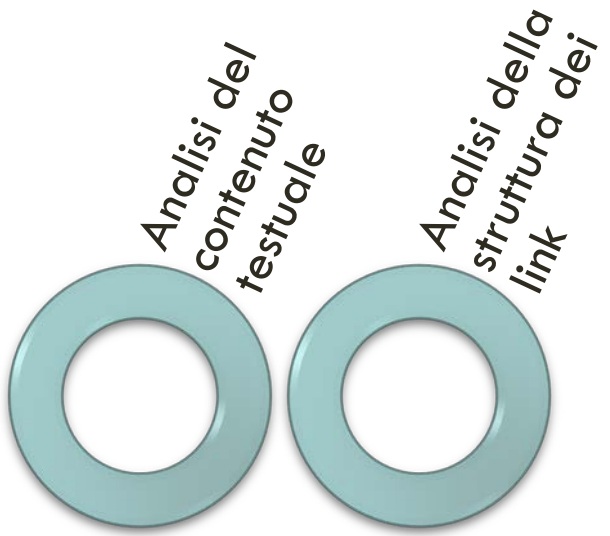


ALGORITMO DI PAGERANK

RANKING: Dato un insieme di pagine **P** e una query **Q** il ranking è definito da una funzione:

$R_q: P \rightarrow \mathbf{R}$ *(insieme dei numeri reali)*

*che associa ad ogni pagina un numero reale che indica la «**RILEVANZA**» di quella pagina nel contesto di quella query.*



ALGORITMO DI PAGERANK

$$\sum_i R_i^{t+1} = \sum_i \left[(1-\alpha) \sum_{j \to i} \frac{R_j^t}{N_j} + \alpha \frac{1}{N} \right] = (1-\alpha) \sum_i \sum_{j \to i} \frac{R_j^t}{N_j} + \alpha$$

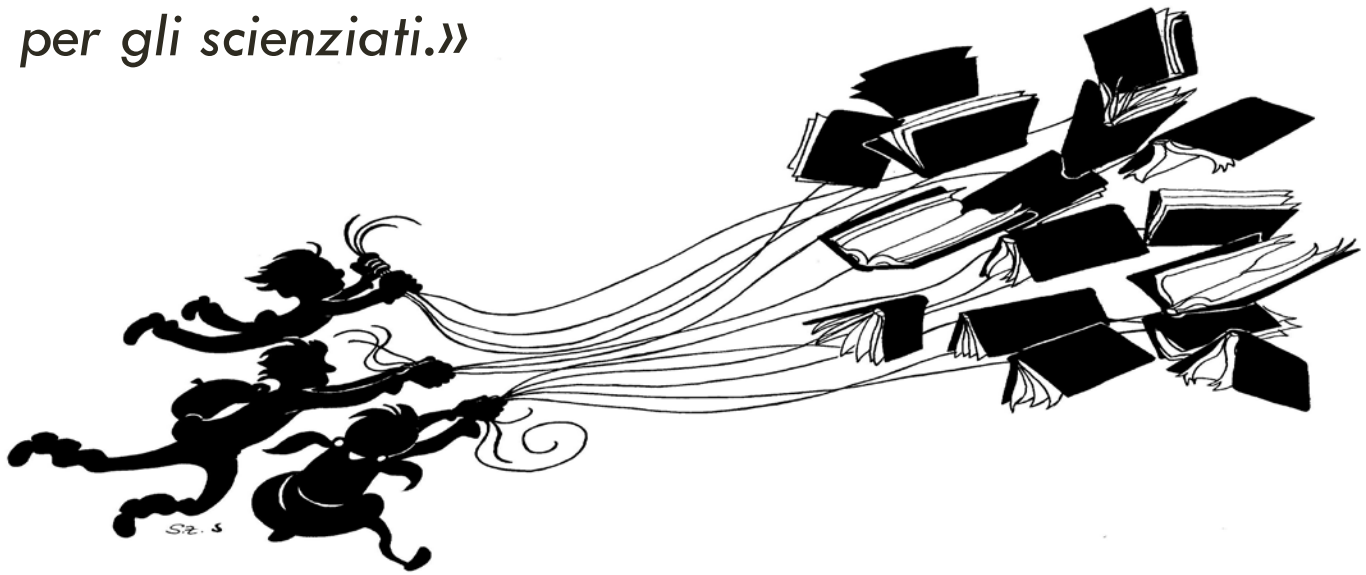
Per ogni nodo j con almeno un arco uscente, il fattore R_j^t/N_j viene sommato per ciascun arco uscente (ce ne sono N_j in tutto):

$$\sum_i R_i^{t+1} = (1-\alpha) \sum_{j \text{ con un arco uscente}} R_j^t + \alpha$$

La storia di Google per capire

https://moz.com/google-algorithm-change

LA RETE E L'INFORMAZIONE

«Di fronte alla giungla di documenti che si aggiungono minuto per minuto, la domanda cruciale è piuttosto la seguente: se lancio un'informazione in rete, qualcuno la noterà? Per esser letti bisogna essere visibili: una banale verità che vale tanto per gli scrittori quanto per gli scienziati.»

Albert-László Barabási in «Link»