Corso di
Web Mining e Retrieval

- Introduzione al Corso -

(a.a. 2009-2010)

Roberto Basili

Overview

- WM&R: Motivazioni e prospettive
- Richiami di Algebra
- Richiami di calcolo delle probabilità
- Introduzione al ML

WM&R: Motivazioni

- *Cos'e' il Web Mining?*
- *Perché IR?*
- *Perché Apprendimento Automatico?*
- *Quale contributo l'IR fornisce alle tecnologie di sfruttamento delle informazioni del Web?*
- *Quali sono le prospettive per l'impiego di tali tecnologie?*

Cos'e' il Web Mining?

- *Web Mining* attualmente si riferisce ad un insieme di tecnologie necessarie allo sfruttamento delle informazioni pubblicamente disponibili nel Web
  - Contenuti: dati ma anche ... persone, luoghi, eventi, concetti, ...
  - Relazioni:
    - Link strutturali
    - Collegamenti tematici, concettuali e interpersonali
    - Ridondanze/analogie
  - Multilingualità
  - Trend e comportamenti collettivi
  - Opinioni

Perché IR?

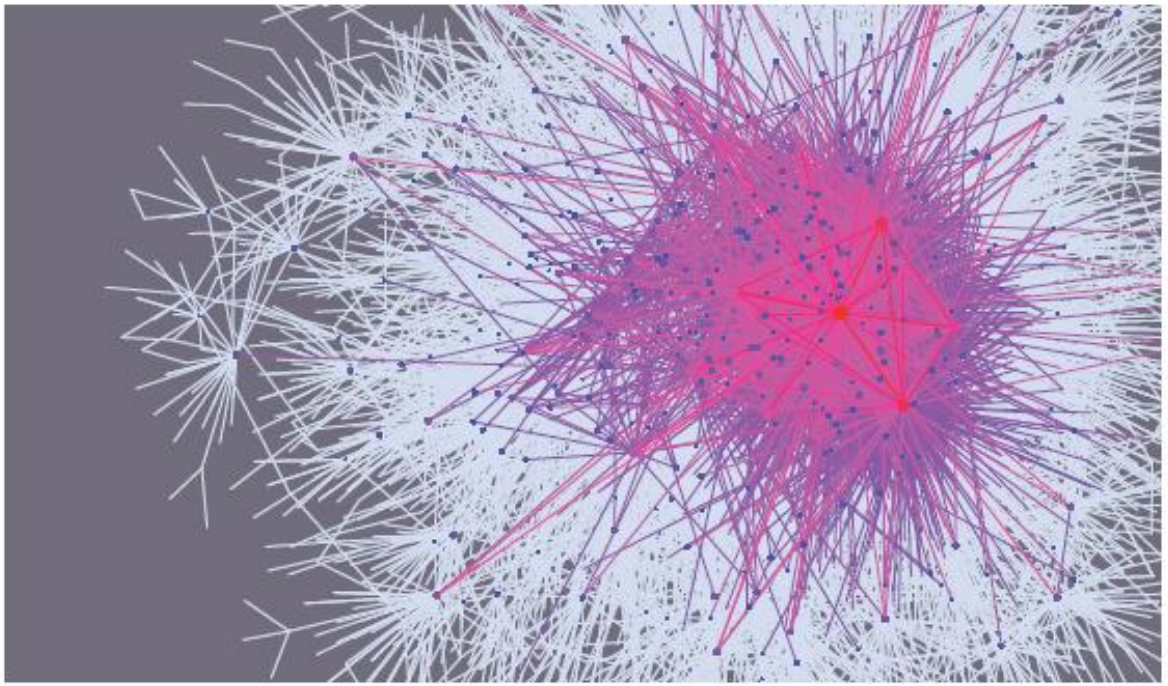
- La taglia delle informazioni in gioco pone il problema della *localizzazione*
- Accedere in modo automatico è possibile solo governando il problema di sapere **dove** si trova una informazione *rilevante*
- La ricerca corrisponde al calcolo di una funzione *aleatoria* di mapping tra requisiti e informazione utile

Machine Learning vs IR?

- La eterogeneità delle informazioni produce significativi effetti di incertezza nel processo di ricerca
  - Incompletezza della informazione
  - Ricchezza di dati, formati e modalità di accesso
  - Requisiti vaghi
  - Aspetti soggettivi
  - Tempestività

ML vs. IR

- La pervasività degli elementi di incertezza rende impraticabile la ricerca di soluzioni esaustive (ottimi globali)
- "*Finding diamonds in the rough*"
  (Fan Chung, UCSD)



ML vs. IR

- Le tecniche di ML propongono una ampia serie di algoritmi, strategie e tecniche per la produzione di soluzioni *sub-ottime* effettive
- Nel processo di *learning* i dati suggeriscono la ipotesi risolutiva per la funzione di *mapping*
- Tale ipotesi migliorerebbe le prestazioni complessive del sistema di base
  - Accuratezza
  - Efficienza computazionale

Machine Learning

- (Langley, 2000): l'Apprendimento Automatico si occupa dei meccanismi attraverso i quali un agente intelligente migliora nel tempo le sue prestazioni *P* nell'effettuare un compito *C*.
- La prova del successo dell'apprendimento è quindi nella capacità di misurare l'incremento ΔP delle prestazioni sulla base delle esperienze *E* che l'agente è in grado di raccogliere durante il suo ciclo di vita.
- La natura dell'apprendimento è quindi tutta nella caratterizzazione delle nozioni qui primitive di *compito*, *prestazione* ed *esperienza*.

Esperienza ed Apprendimento

- L'esperienza, per esempio, nel gioco degli scacchi può essere interpretata in diversi modi:
  - i dati sulle vittorie (e sconfitte) pregresse per valutare la bontà (o la inadeguatezza) di strategie e mosse eseguite rispetto all'avversario.
  - valutazione fornita sulle mosse da un docente esterno (oracolo, guida).
  - Adeguatezza dei comportamenti derivata dalla auto-osservazione, cioè dalla capacità di analizzare partite dell'agente contro se stesso secondo un modello esplicito del processo (partita) e della sua evoluzione (comportamento, vantaggi, ...).

Esperienza ed Apprendimento (2)

- Possiamo quindi parlare nei tre casi di:
  - *apprendimento per esperienza*, o *induttivo*, (partite eseguite e valutate in base al loro successo finale)
  - *apprendimento supervisionato* (cioè partite, strategie e mosse giudicate in base all'oracolo)
  - *apprendimento basato sulla conoscenza* relativa al task, che guida la formazione di modelli del processo e modelli di comportamento adeguato.

Apprendimento senza supervisione

- In assenza di un oracolo o di conoscenze sul task esistono ancora molti modi di migliorare le proprie prestazioni, ad es.
  - Migliorando il proprio modello del mondo (acquisizione/*discovery* della conoscenza)
  - Migliorando le proprie prestazioni computazionali (ottimizzazione)

Apprendimento senza supervisione

- E: Al termine del processo di acquisizione il sistema
  - dispone di un sistema di classi e relazioni indotte che migliora la sua interazione futura con l'ambiente operativo (ad es. l'utente)
  - Il miglioramento avviene quindi almeno rispetto agli algoritmi di ricerca: la organizzazione gerarchica consente di esaminare solo i membri dell'insieme in alcune classi (i generi).

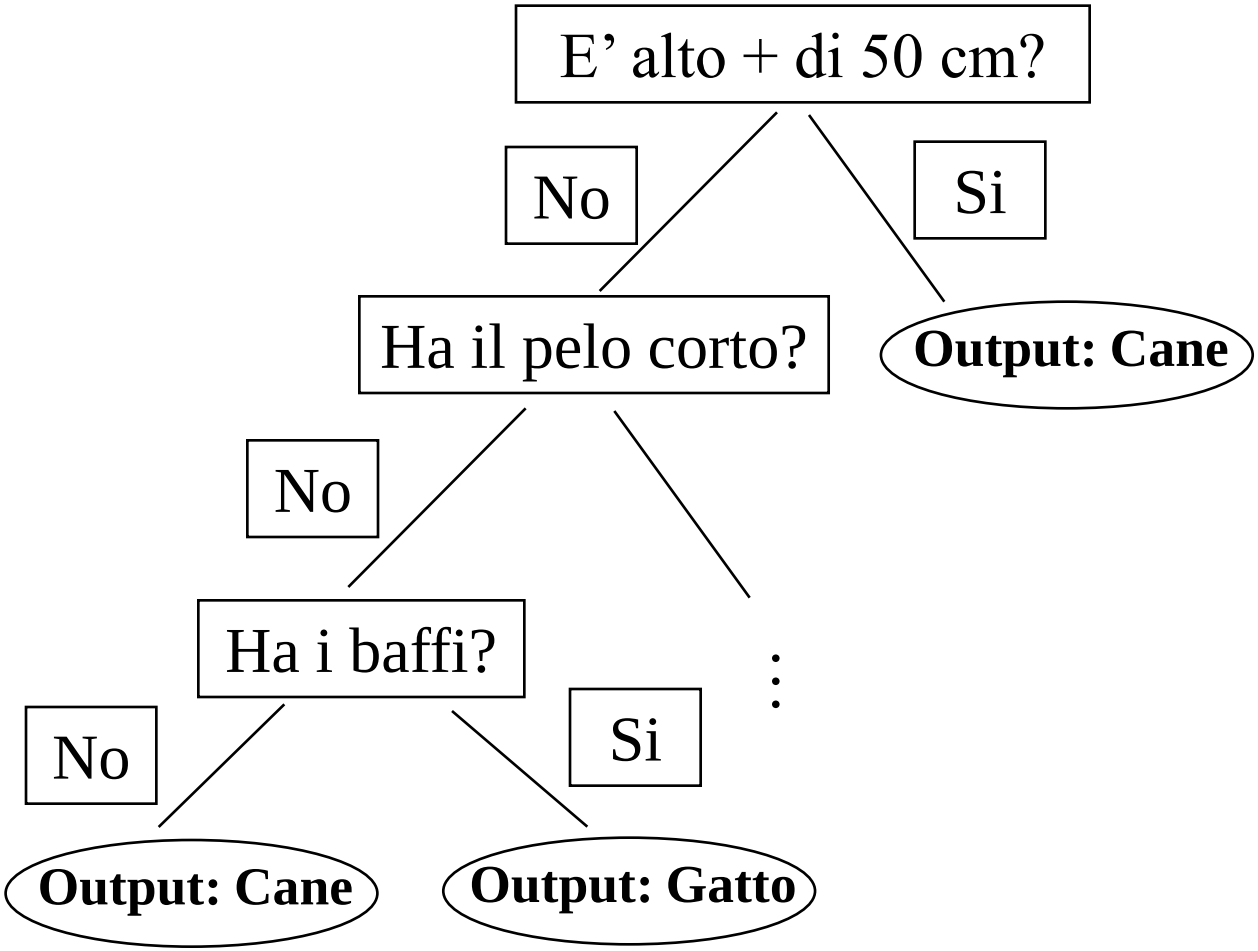
L'apprendimento automatico

- Apprendere una funzione da esempi:
  - a valori reali, *regressione*
  - a valori interi finiti, *classificazione*
- Supponiamo di volere apprendere una funzione intera:
  - 2 classi, *gatto* e *cane*
  - $f(x) \rightarrow \{gatto, cane\}$
- Dato un insieme di esempi per le due classi
  - Si estraggono le features (*altezza*, *baffi*, *tipo di dentatura*, *numero di zampe*).
- Si applica l'algoritmo di learning per generare *f*

Algoritmi di Apprendimento

- Funzioni logiche booleane, (ad es., alberi di decisione).
- Funzione di Probabilità, (ad es., classificatore Bayesiano).
- Funzioni di separazione in spazi vettoriali
  - Non lineari: KNN, reti neurali multi-strato,...
  - **Lineari**, percettroni, **Support Vector Machines**,...
- Trasformazioni di spazi: embeddings, analisi spettrale

Alberi di decisione (Gatti/Cani)



Web IR?

- I processi di IR studiati in domini antecedenti all'affermarsi del Web devono essere estesi ed adattati rispetto alla maggiore ricchezza ed ai problemi maggiori che tali scenari presentano
  - Complessità strutturale: contenuti, topologia e uso
  - Affidabilità dell'informazione
  - Multimodalità, Multimedialità
  - Partecipazione (aspetti sociali)

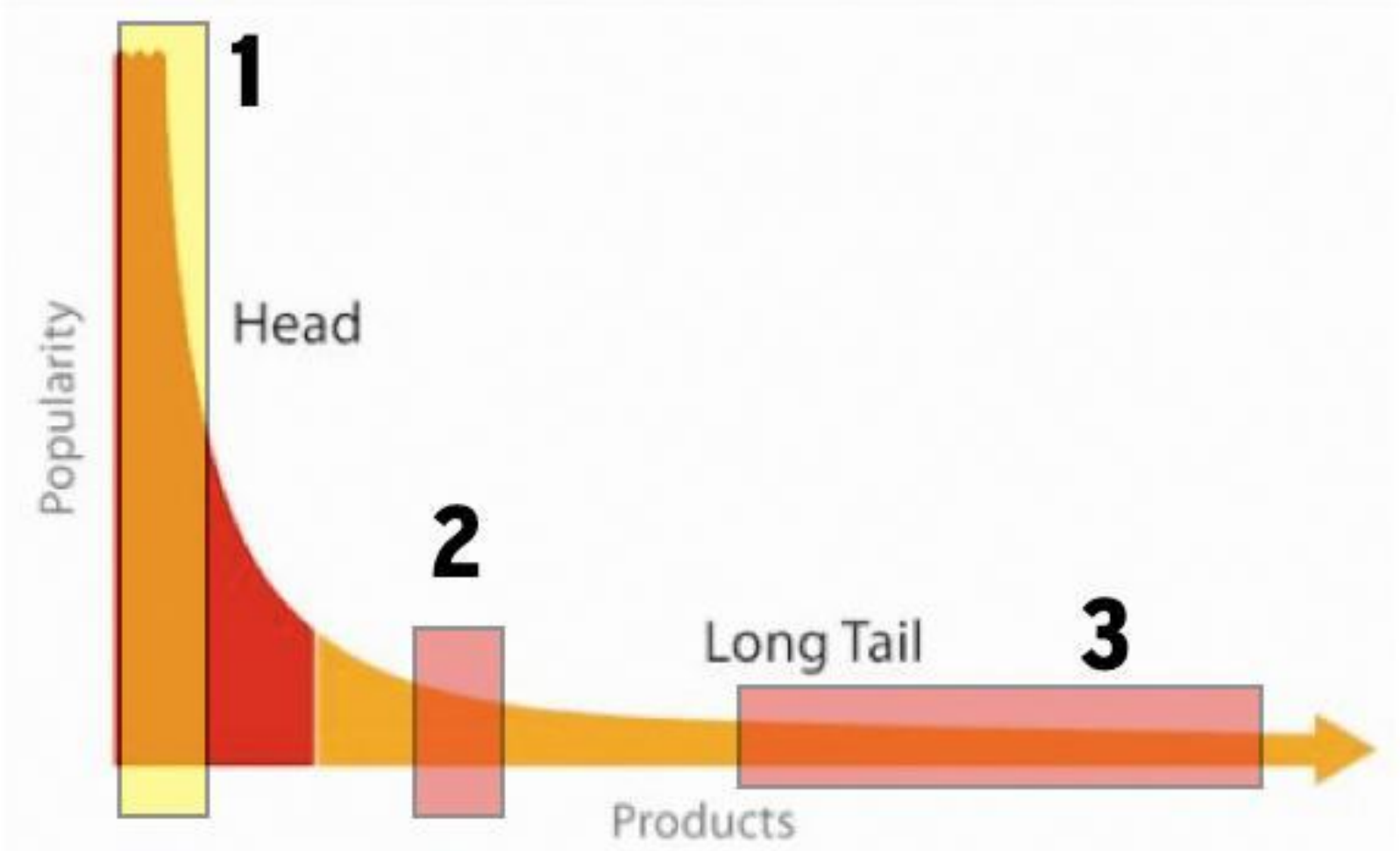
Web IR

- Processare dati Web: trovare contenuti, trovare link, ...
- Web Crawling
- Web Search: indici, link analysis
- Classificazione: contenuti pesati, collegamenti e formati, autorità, temporizzazione
- Meta-ricerca
- Analisi dei Link

Prospettive delle tecnologie WM&R

- Crescita esponenziale della taglia dei problemi
- Crescente interesse verso processi di IR agenti su dati complessi (multimediali, sociali)
- Web partecipativo: Web 2.0
- Ruolo crescente della mediazione degli strumenti informatici
  - Software come Servizio
  - Personalizzazione

La lunga Coda



Web Sociale

2008 U.S. Social Network Usage (Comscore)

	12/1/2007 (millions)	12/1/2008 (millions)	Yearly Growth	Monthly Growth
MySpace	69	76	10%	0.8%
Facebook	35	55	57%	3.8%
Classmates	10	16.6	66%	4.3%
LinkedIn	2.9	6.3	117%	6.7%
Bebo	NA	4.9		
Ning	0.8	3.9	388%	14.1%
Friendster	1.8	1.7	-6%	-0.5%

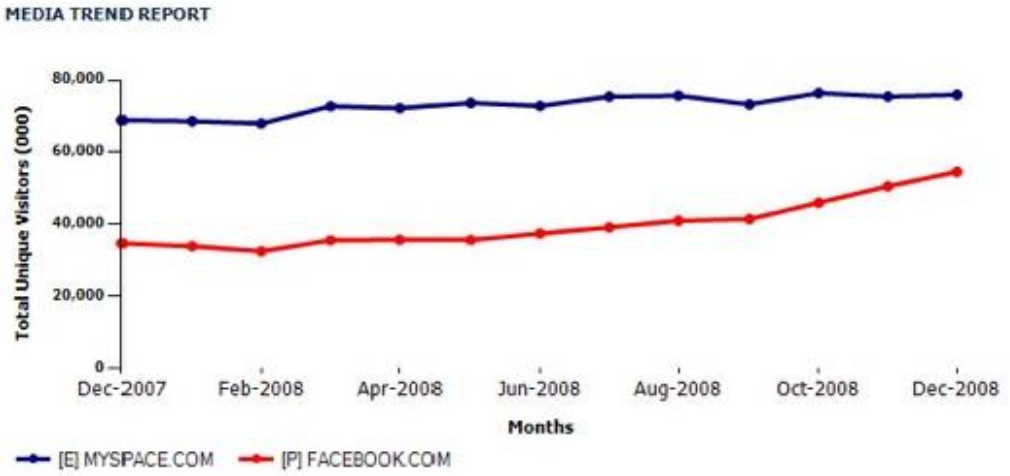
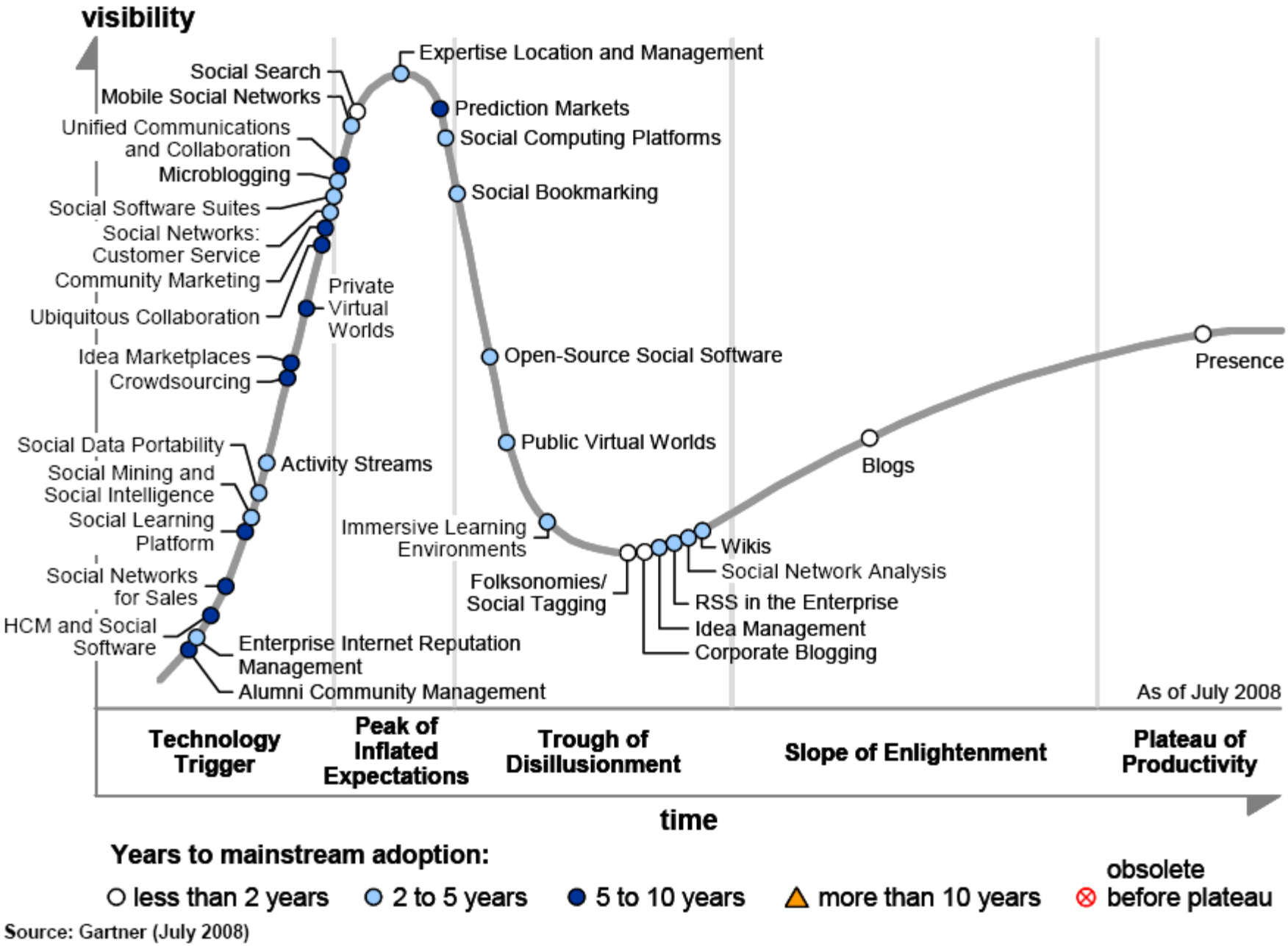
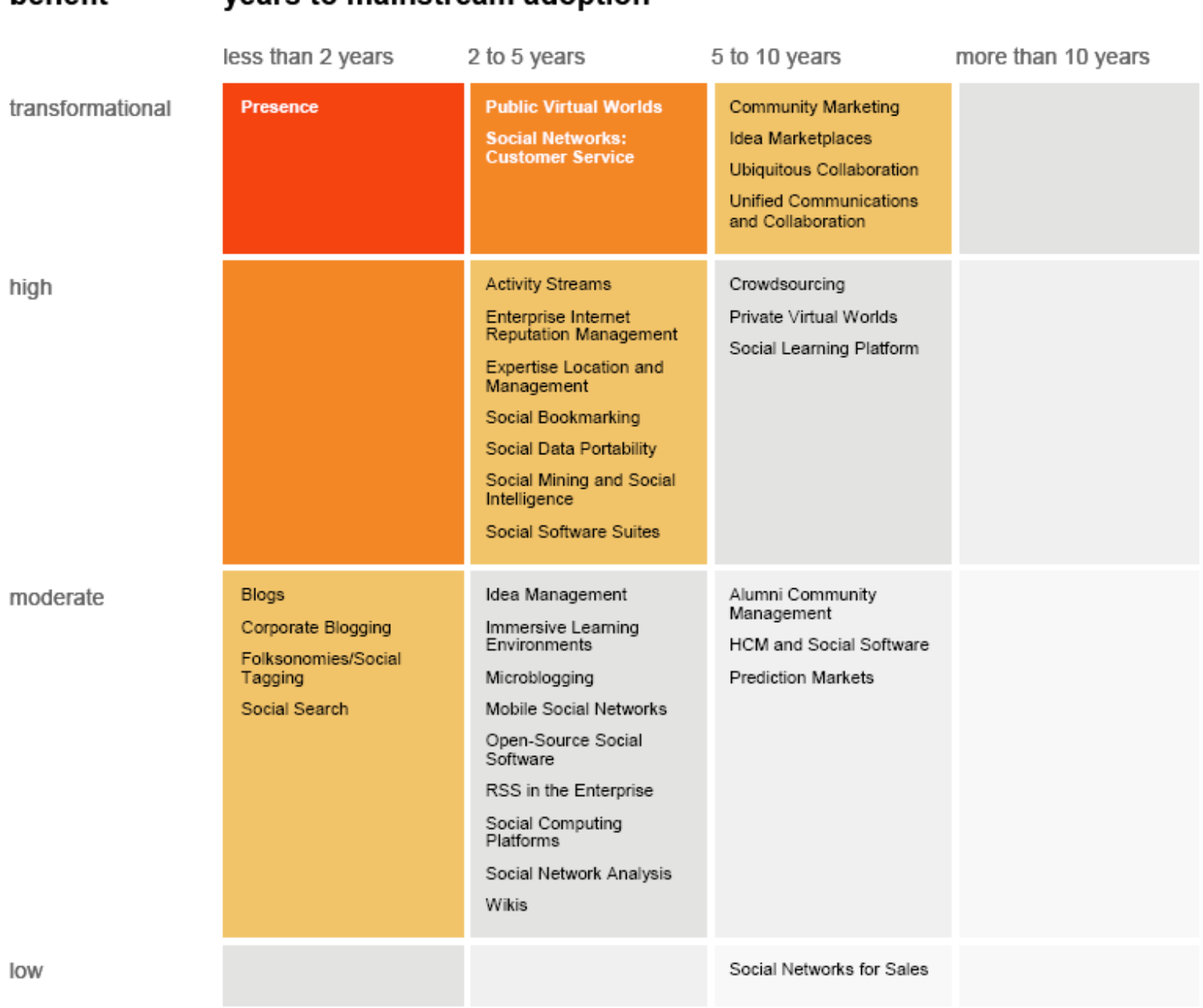


Figure 1. Hype Cycle for Social Software, 2008



Source: Gartner (July 2008)

Figure 2. Priority Matrix for Social Software, 2008



As of July 2008

Source: Gartner (July 2008)