Data Mining
a.a. 2010-2011

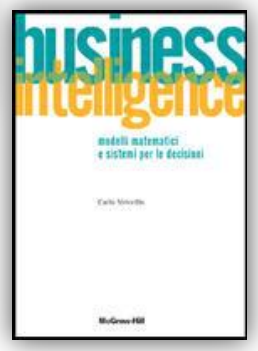
- Docente: Mario Guarracino
  - mario.guarracino@cnr.it
  - tel. 081 6139519
  - http://www.na.icar.cnr.it/~mariog

Informazioni logistiche

- Orario delle lezioni
  - A partire dall' 19.10.2010, Martedì h: 09.50 – 16.00 Aula 2 - SAN BENEDETTO
- Ricevimento
  - Alla fine delle lezioni, per appuntamento (e-mail, telefono,…)
- Organizzazione delle lezioni
  - Lezioni frontali ed in laboratorio

Informazioni generali

- Libro di testo
  - Paolo Giudici, *Data Mining*, McGraw-Hill, 2005
- Altri riferimenti
  - Carlo Vercellis, *Business intelligence*, McGraw-Hill, 2006. ☺



- Materiale didattico
  - lucidi delle lezioni disponibili sul sito del corso
  - …

Informazioni generali

- Iscrizione al corso
  - invio di una e-mail all'indirizzo del docente (preferibilmente da un indirizzo di posta dell'università)
    - Subject: **Iscrizione DM2010**
- Modalità d'esame
  - E' previsto un progetto e un orale
  - Contribuiscono alla valutazione:
    - la partecipazione attiva al corso
    - Il progetto
    - la prova orale

Prerequisiti

- I contenuti di
  - Sistemi informatici orientati ai servizi in rete per le PP.AA .
- Non è prevista alcuna propedeuticità formale

Obiettivi

Obiettivo del corso è di illustrare i *processi di analisi delle basi di dati*, orientati a produrre risultati utili per le decisioni.

Lo scopo è di comprendere la *struttura* e le *funzioni* dei sistemi informativi mediante lo studio di *algoritmi, metodi e strumenti* e la loro implementazione in sistemi reali.

Partendo dai processi decisionali, verranno illustrati gli strumenti di data warehouse e i metodi di data mining.

Si illustreranno infine casi concreti di applicazione.

Come posso partecipare?

- Prendendo parte alle lezioni ed alle discussioni,
- Arricchendo il materiale del corso:
  - FAQ,
  - bibliografia,
  - URL,
  - soluzioni agli esercizi,
  - …
- Tesi, tesine e progetti,
- …

Programma

- Argomenti del corso
  - Introduzione al data mining
  - Data mining e statistica
  - Organizzazione dei dati
  - Analisi esplorativa dei dati
  - Metodi computazionali per il data mining
  - Modelli statistici per il data mining
  - Casi di studio

Perché?

- La “borsa degli strumenti”.
- Conoscere a fondo lo strumento che si utilizza permette di ottenere risultati migliori.
- Estrarre conoscenza utile da ingenti moli di dati, è la chiave del successo dei decision maker nella pubblica amministrazione e nelle imprese.
- Anche i forni a microonde prendono decisioni a partire dall’analisi dei dati!
- “Tu sei esperto di *scienze e tecniche delle amministrazioni pubbliche*, giusto?!”

Data Mining

- L'avvento di tecnologie di memorizzazione a basso costo e la diffusione della connettività hanno reso più agevole l'accesso a grandi quantità di dati.
- I dati disponibili sono eterogenei per origine, contenuto e rappresentazione.
  - Transazioni commerciali, finanziarie, amministrative;
  - Percorsi di navigazione web, email, ipertesti;
  - Test clinici,…
- La loro presenza apre scenari e opportunità prima impensabili.
- Per *data mining* (DM) intenderemo l'insieme delle metodologie e modelli che esplorano i dati per ricavarne informazioni e quindi conoscenza.

Quali problemi possiamo risolvere?

Esempio 1

- Un operatore di telefonia mobile nota un aumento nel numero delle disattivazioni tra i propri clienti.

- Ha a disposizione un budget per *customer retention* per 200 mila tra i 2 milioni clienti.

- Come può procedere nella scelta dei destinatari della promozione?

Quali problemi possiamo risolvere?

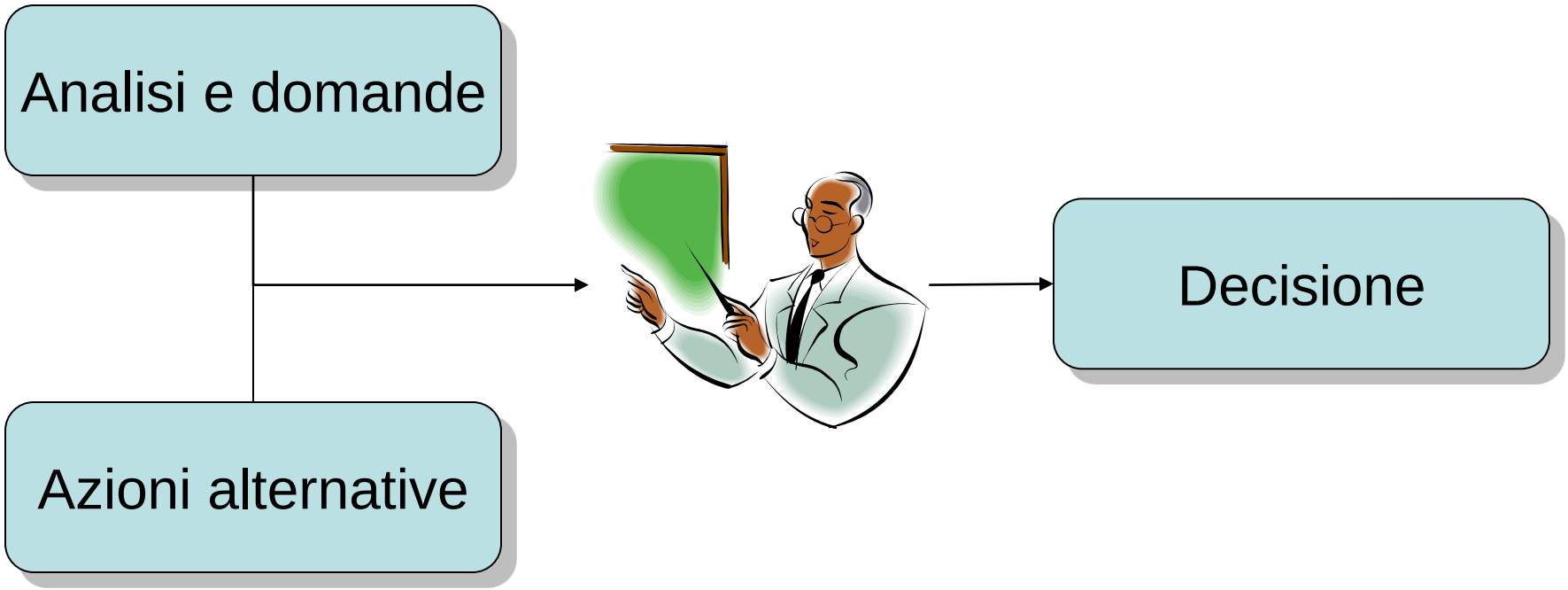
Esempio 2

- Un'azienda vuole ottimizzare i costi logistici e produttivi.
- Ha una decina di stabilimenti che devono approvvigionarsi, produrre e distribuire secondo le esigenze del mercato, che variano durante l'anno.
- Come si può sviluppare un piano logistico ottimale?

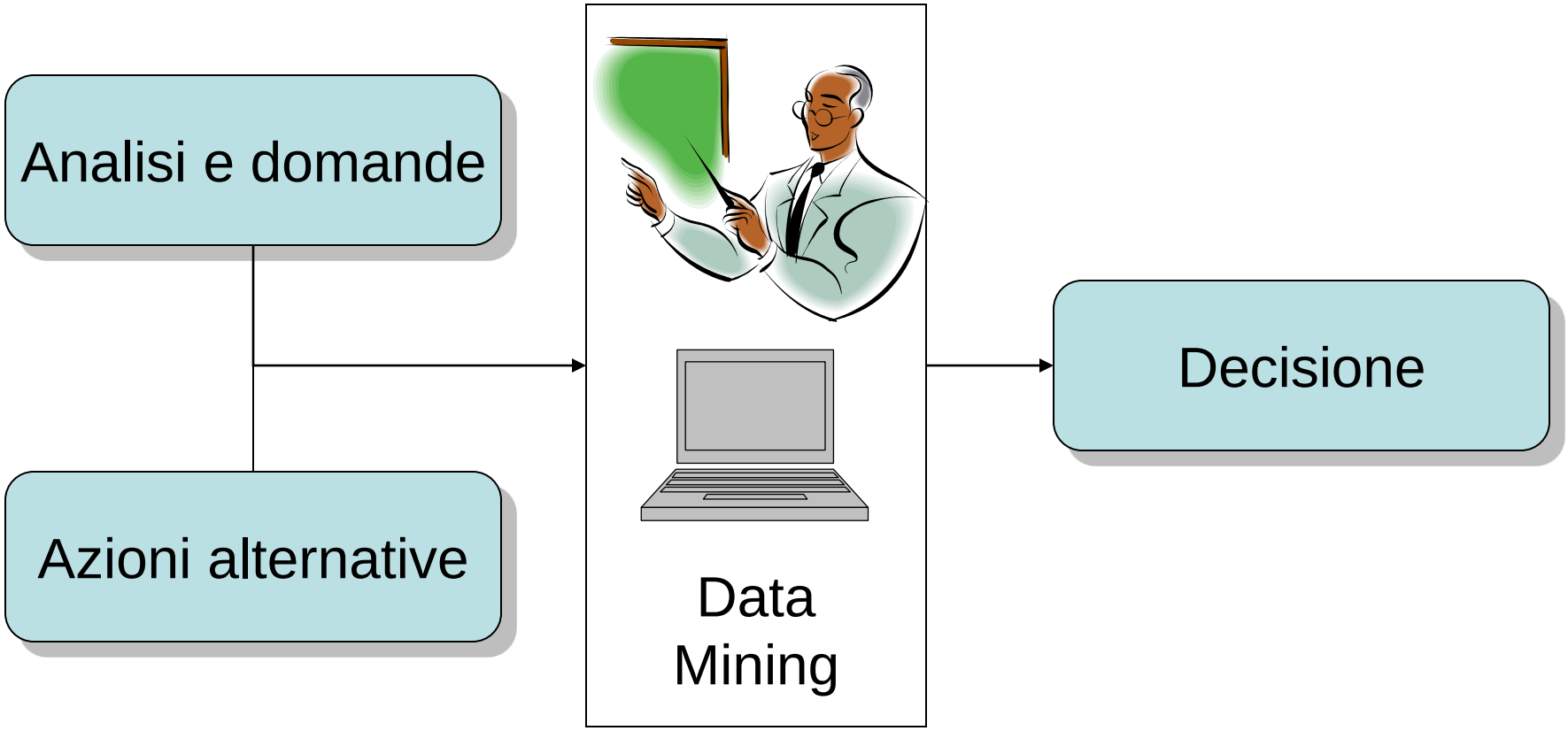
Decisioni efficaci e tempestive

- La disponibilità di informazioni e conoscenze ricavate da analisi quantitative permette di prendere decisioni efficaci.
- La capacità di reagire dinamicamente alle azioni dei competitori e alle esigenze del mercato rappresenta un fattore decisivo di successo.
- E' necessario quindi avere a disposizione strumenti e metodologie che permettono di individuare decisioni efficaci e tempestive.

Vantaggi del DM



Vantaggi della DM



- Più alternative analizzate
- Conclusioni più precise
- Decisioni efficaci e tempestive

Dati, informazioni e conoscenza

- I dati di natura amministrativa, logistica e commerciale delle imprese e della pubblica amministrazione sono, per natura, eterogenei.

- Anche se raccolti in modo sistematico e strutturato, tali dati non sono direttamente utilizzabili nell'ambito dei processi decisionali.

- E' necessario organizzarli ed elaborarli mediante opportuni strumenti che li trasformino in informazioni e conoscenze applicabili dai *decision maker*.

Dati, informazioni e conoscenza

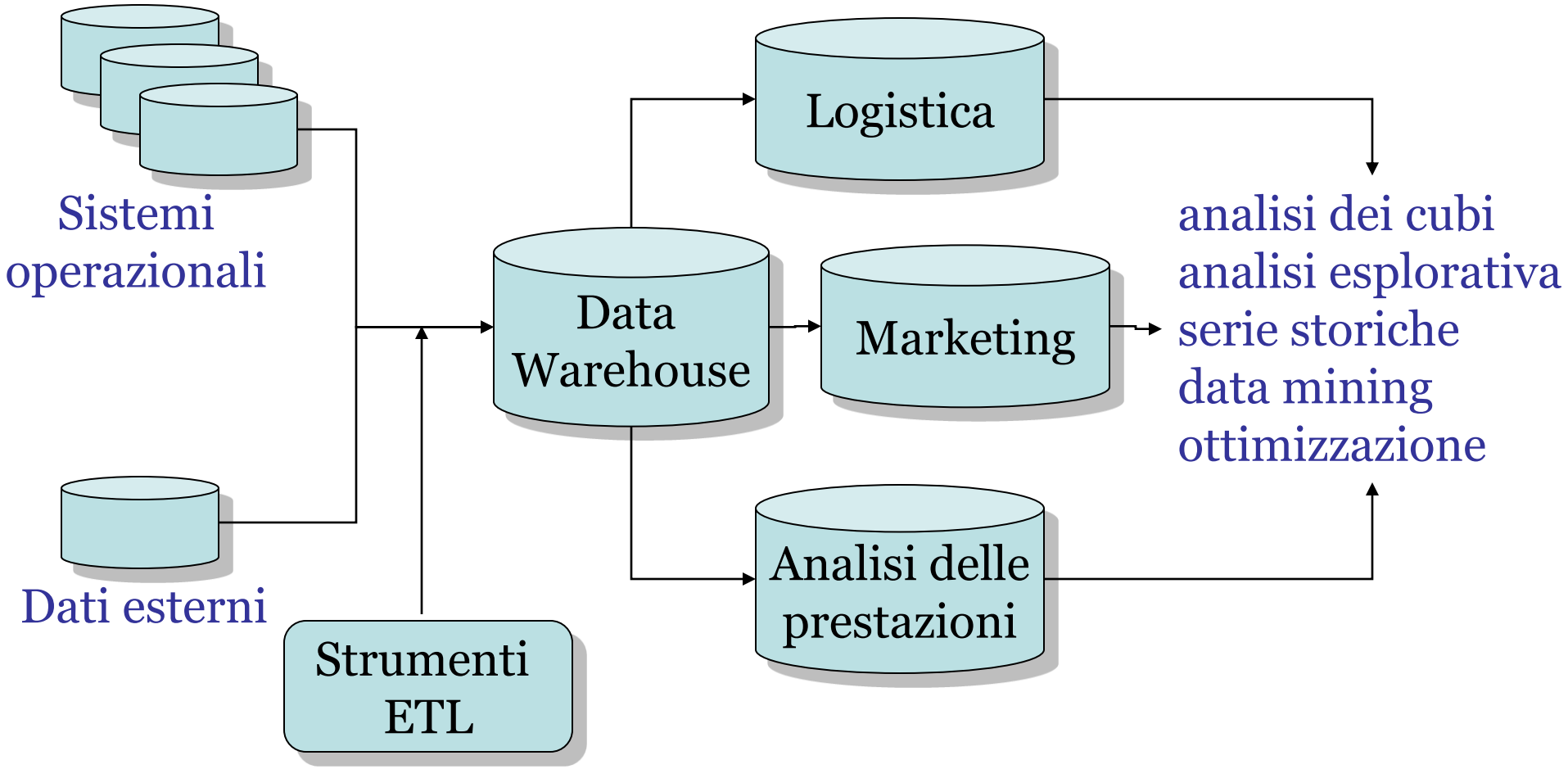
- **Dati**: Codifica strutturata delle singole entità primarie e delle transazioni che coinvolgono due o più entità primarie.
  - **Esempio**: Base di dati dei clienti di un supermercato.
- **Informazioni**: Risultato di operazioni di estrazione e elaborazione compiute a partire dai dati.
  - **Esempio**: Clienti che hanno ridotto di più del 50% l'importo mensile d'acquisto negli ultimi tre mesi.
- **Conoscenza**: Informazioni contestualizzate e arricchite dall'esperienza e dalle competenze del decision maker.
  - **Esempio**: Analisi delle vendite e del contesto territoriale.

Ruolo dei modelli matematici

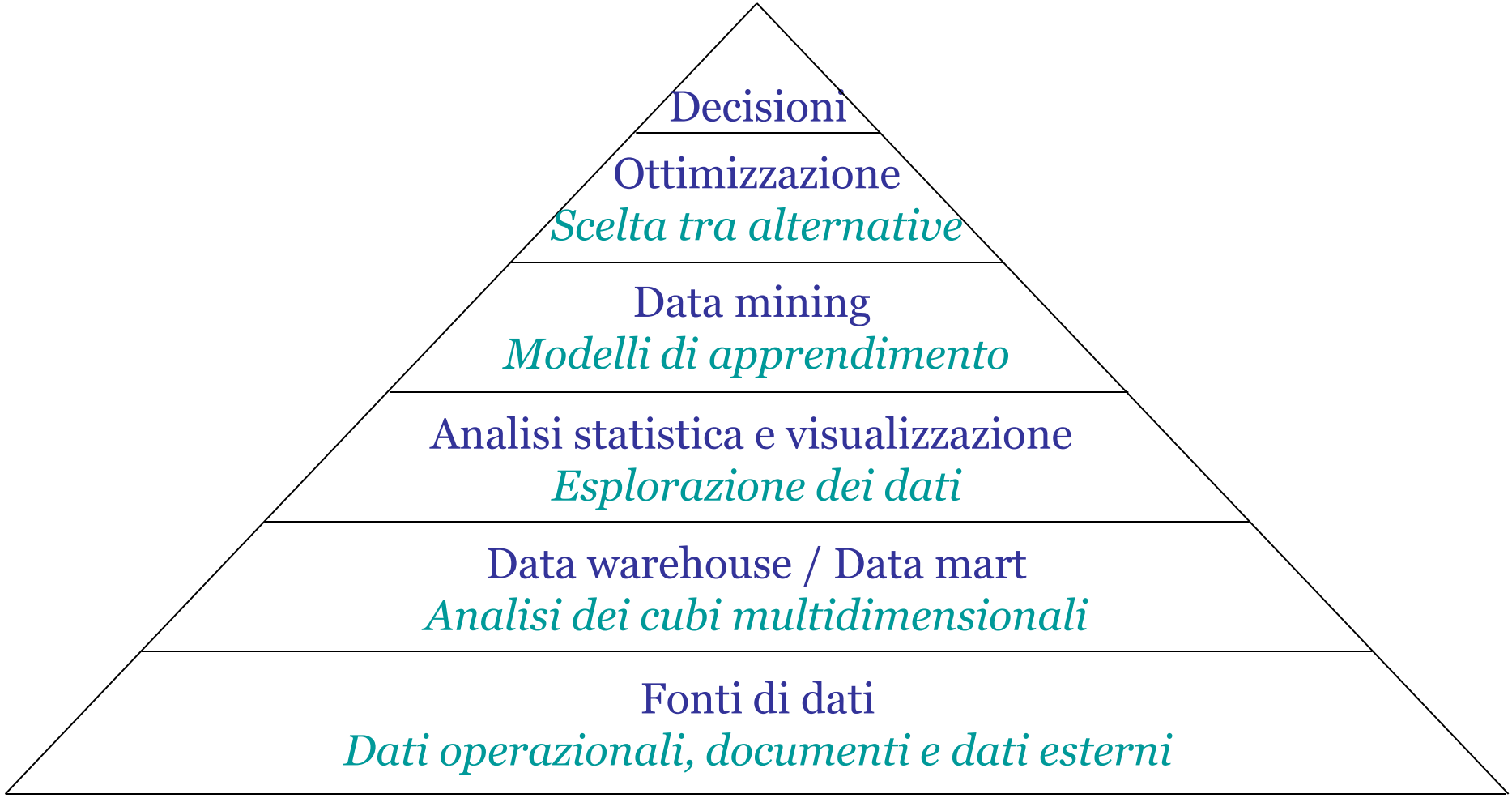
- Il data mining offre al decision maker informazioni e le conoscenze ricavate dai dati mediante opportuni modelli matematici.

- Questo tipo di analisi tendono a promuovere un orientamento scientifico e razionale nella gestione delle imprese e della pubblica amministrazione:
  - Individuare gli obiettivi delle analisi e degli indicatori di prestazioni,
  - Sviluppare modelli matematici che relazionano le variabili di controllo con i parametri e le metriche di valutazione,
  - Analizzare gli effetti sulle prestazioni delle variazioni delle variabili di controllo.

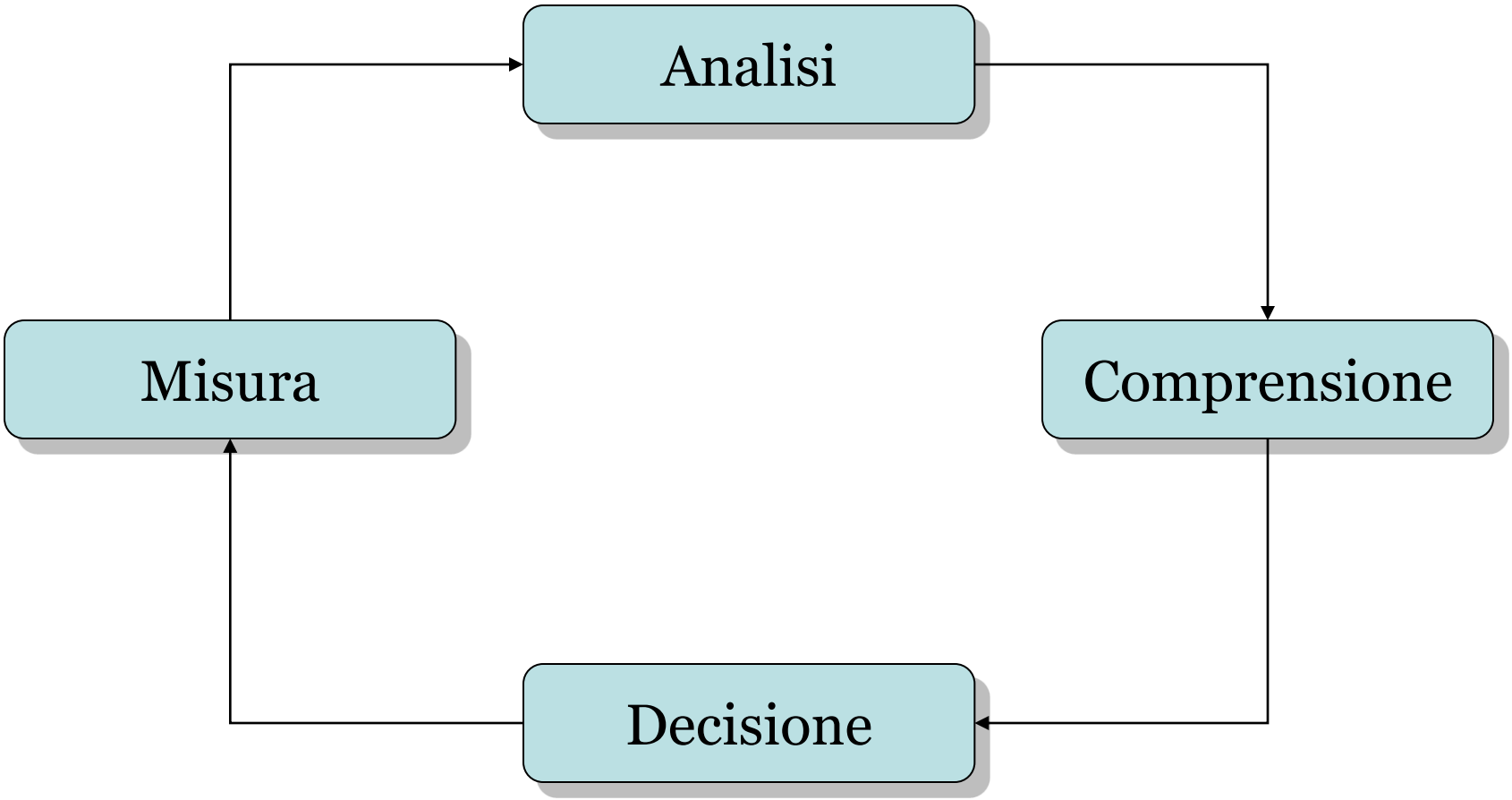
Architettura di business intelligence



Componenti di un ambiente BI



Fasi di analisi BI

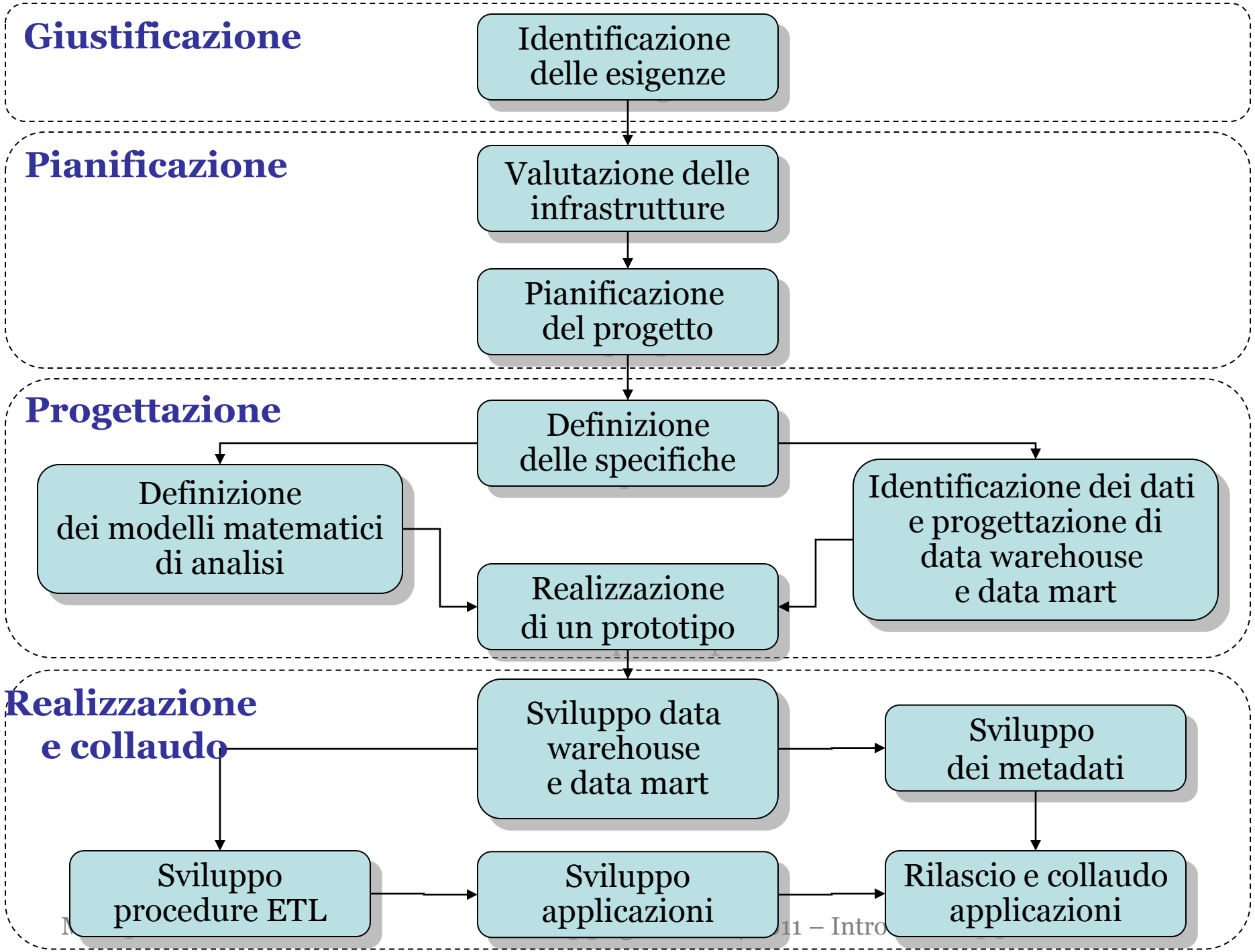


Fattori abilitanti

- **Tecnologie** Le tecnologie hardware e software, disponibili ovunque e a basso prezzo, ha permesso di derivare ed utilizzare sofisticati algoritmi di calcolo.

- **Metodologie analitiche** La rappresentazione visuale dei dati non è sufficiente ad attivare un processo attivo di analisi

- **Risorse umane** la capacità dei knowledge worker rappresenta il patrimonio principale di ciascuna organizzazione.


Giustificazione
Identificazione delle esigenze
Pianificazione
Valutazione delle infrastrutture
Pianificazione del progetto
Progettazione
Definizione delle specifiche
Definizione dei modelli matematici di analisi
Identificazione dei dati e progettazione di data warehouse e data mart
Realizzazione di un prototipo
Realizzazione e collaudo
Sviluppo data warehouse e data mart
Sviluppo dei metadati
Sviluppo procedure ETL
Sviluppo applicazioni
Rilascio e collaudo applicazioni

Sommario

- Abbiamo visto:
  - Perché è interessante studiare il data mining;
  - Quali problemi si possono risolvere;
  - La differenza tra *dati, informazioni e conoscenza*;
  - A cosa servono i *modelli matematici* in questo contesto;
  - Come sono logicamente organizzate le *architetture* di BI;

Nella prossima lezione

- Sistemi di supporto alle decisioni:
  - Rappresentazione dei processi decisionali;
  - Evoluzione dei sistemi informativi;
  - Definizioni di DSS;
  - Sviluppo dei DSS;