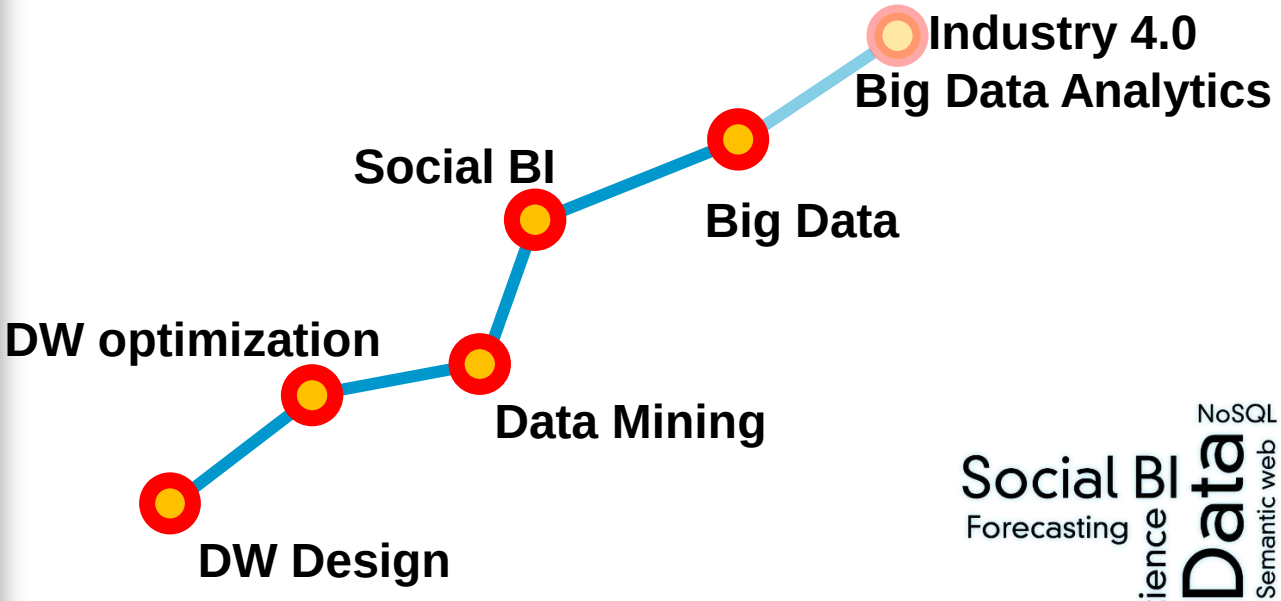
Data Mining (Intelligenza Artificiale e Machine Learning)

Matteo Francia

BBS
BOLOGNA BUSINESS SCHOOL

Il Business Intelligence Group

Il Business Intelligence Group svole ricerche su analisi, tecniche e tecnologie nell'ambito della data analysis



BIG Expertise

European funding

- *PANDA* (pattern management in DM)
- *ENPADASI* (EU Nutritional Phenotype Assessment and Data Sharing Initiative)
- *TOREADOR* (As-a-service Big Data Analytics)
- *WeLaser* (Laser-based Robotic Weeding)

Public funding

- *D2I* (integration and mining of heterogeneous DBs)
- *WISDOM* (ontology-enhanced web searching)
- *WebPolEU*(Comparing Social Media and Political Participation across EU)
- *GenData2020* (data-centric genomic computing)
- *DyNamiTE* (Digital fightiNg Tax Evasion)
- *MO.RE.Farming* (Big Data for Precision Farming)
- *INNOFRUVE* (Ricerca industriale ed innovazione nel comparto ortofrutta)
- *AgroBigDataScience* (Big Data for Precision Farming)

Private funding (2015-2021)

- *Data Mining in the Fashion Field* with Valentino
- *Set-up of a Social Business Intelligence framework* with Amadori s.p.a.
- *Feasibility study for a Social Business Intelligence system* with DOXA
- *Anomaly detection in the gas network* with HERA spa
- *Harnessing Wellness Knowledge* with Technogym
- *Methodological and Scientific Support to several Public bodies* With Ministry of Justice,Economy and Fin
- *Vaccine monitoring* with Regione Veneto & ONIT
- *Intelligent Monitoring Systems for Critical Environments* with Leonardo-Finmeccanica
- *Data-driven budgetting* with Teddy
- *Digital Transformation* with BRT, PLT Energia

Acknowledgments

This is the result of a joint work

- DISI @ UniBO --- Department of Computer Science and Engineering
  - **Prof. Matteo Golfarelli**
  - Prof. Matteo Francia
  - Prof. Enrico Gallinucci
  - Dr. Chiara Forresi
  - Dr. Joseph Giovanelli

A mandatory premise: a module with multiple levels of understanding

- Talking about technical topics to
  - ✓ a non-technical audience is hard and sometimes frustrating
  - ✓ a heterogeneous background audience is even harder and often frustrating
- … listening is typically worse!
- I will avoid unnecessary technicalities but… sometimes they are necessary!
- Don't be afraid of technicalities
  - ✓ If something is not clear but you believe can be useful to your profile, please ask!
  - ✓ If something is not clear and useless to your profile, focus on the whole picture

AI, Machine Learning & Data Mining

Il termine **artificial intelligence** (AI) indica la parziale riproduzione delle capacità intellettuali umane (in particolare quelle di apprendimento del riconoscimento e della scelta) tramite la definizione di modelli matematici e, più nello specifico, tramite lo sviluppo di machine controllate dal computer.

Si tratta di disciplina molto ampia che presenta molte sotto-aree:

- Trial and Error Search, Heuristics, Evolutionary computing
- Knowledge Representation and Reasoning
- Automated Theorem Proving
- Expert Systems
- Planning, Coordination and Manipulation
- Intelligent Agents
- Robotics
- Automatic Programming
- Natural Language Processing
- Vision and Speech
- Machine Learning

AI, Machine Learning & Data Mining

Il **Machine Learning** è considerato uno degli approcci più importanti dell'AI poiché imparare è una capacità chiave e permette di migliorare ed evolvere.

Imparare un comportamento dai dati/esempi forniti semplifica lo sviluppo di applicazioni e consente di gestire la complessità delle applicazioni del mondo reale, a volte troppo complesse per essere modellate in modo efficace.

Machine learning

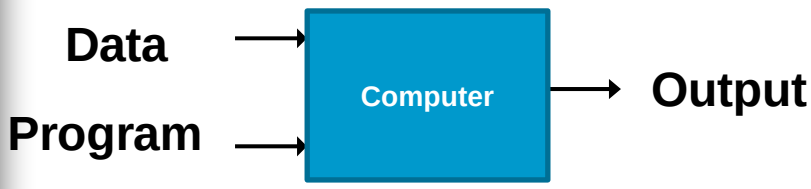
- Machine Learning is the science (and art) of programming computers so they can learn from data
  - ✓ “Learning is any process by which a system improves performance from experience.” - Herbert Simon
  - ✓ Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed — Arthur Samuel, 1959
  - ✓ A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E — Tom Mitchell, 1997

What is Machine Learning?

Definition by Tom Mitchell (1998):

- Machine Learning is the study of algorithms that
  - ✓ Improve their performance P
  - ✓ At some task T
  - ✓ With experience E
- A well-defined learning task is given by <P, T, E>

- Traditional Programming



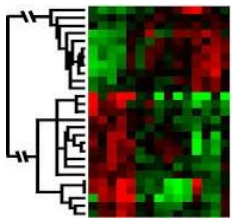
- Machine Learning

When Do We Use Machine Learning?

- ML is used when:
  - ✓ Human expertise does not exist (navigating on Mars)
  - ✓ Humans can't explain their expertise (speech recognition)
  - ✓ Models must be customized (personalized medicine)
  - ✓ Models are based on huge amounts of data (genomics)

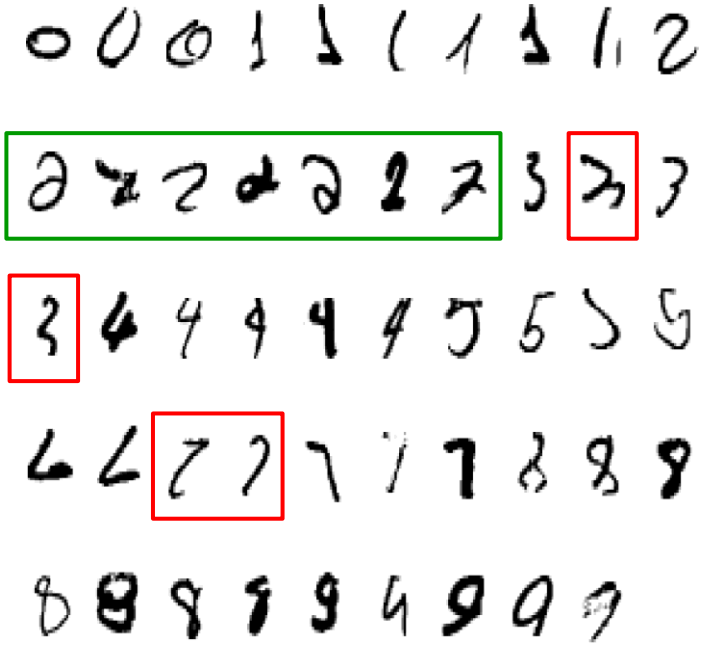
- Learning isn't always useful:
  - ✓ There is no need to "learn" to calculate payroll



When Do We Use Machine Learning?

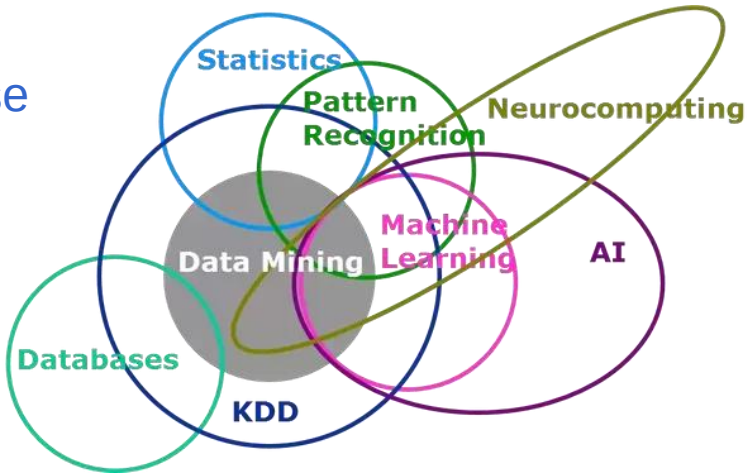
- A classic example of a task that requires machine learning:
  - ✓ It is hard to say what makes a 2



AI, Machine Learning & Data Mining

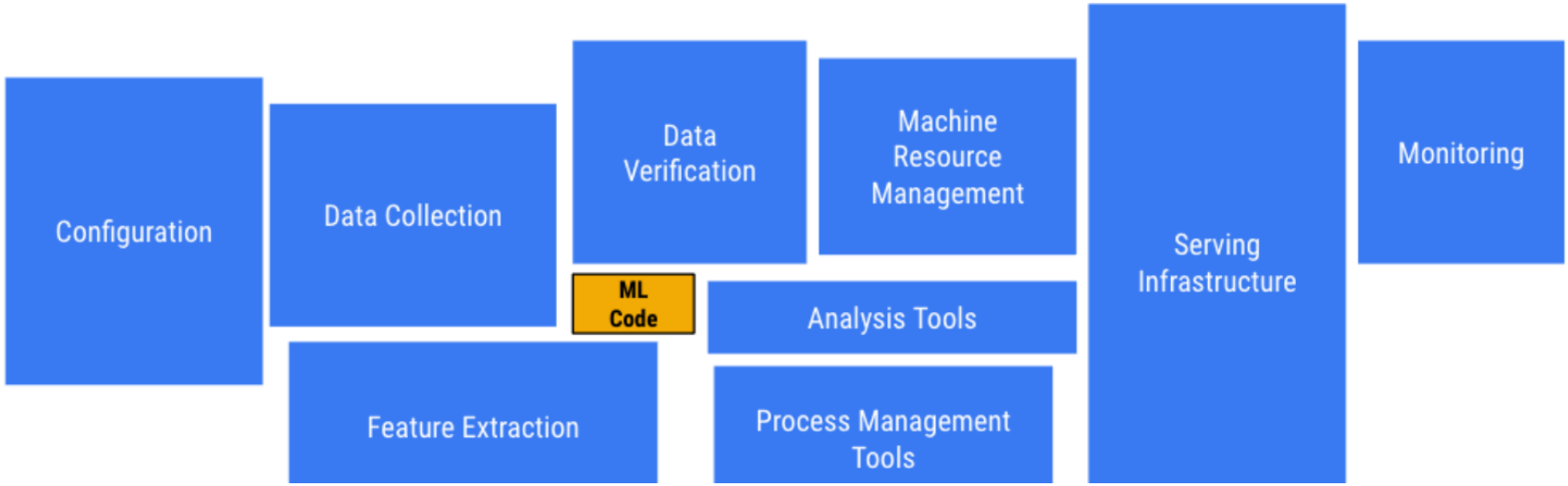
- Sebbene fortemente interrelati tra loro, il termine machine learning è formalmente distinto dal termine Data Mining con il quale si indica il processo computazionale di scoperta di pattern in grandi dataset utilizzando metodi di machine learning, intelligenza artificiale, statistica e basi di dati.

- A parte la fase di analisi vera e propria, il data mining copre aspetti di:
  - ✓ Gestione del dato e pre-processing
  - ✓ Modellazione
  - ✓ Identificazione di metriche di interesse
  - ✓ Visualizzazione



AI, Machine Learning & Data Mining

- Il ruolo del Machine Learning in un progetto reale di data mining è reso bene dalla seguente immagine che elenca le attività necessarie. A rettangoli più grandi corrispondono alle attività a cui è dedicato più tempo.



Analytics

- Il termine **analytics** si riferisce al software utilizzato per la scoperta, comprensione e condivisione di modelli rilevanti nei dati. Le analisi si basano sull'uso simultaneo di statistiche, apprendimento automatico e tecniche di ricerca operativa. Le analisi spesso sfruttano tecniche di visualizzazione avanzate
- Analytics nella BI 2.0 gioca lo stesso ruolo giocato dal data mining nella BI 1.0
- Le soluzioni di data mining si sono diffuse molto meno di quelle DW a causa di:
  - ✓ Complessità e costi
  - ✓ Necessità di un esperto per la comprensione dei risultati
  - ✓ Mancanza di certezza nel raggiungimento degli obiettivi del progetto

TED TALKS sul Machine Learning

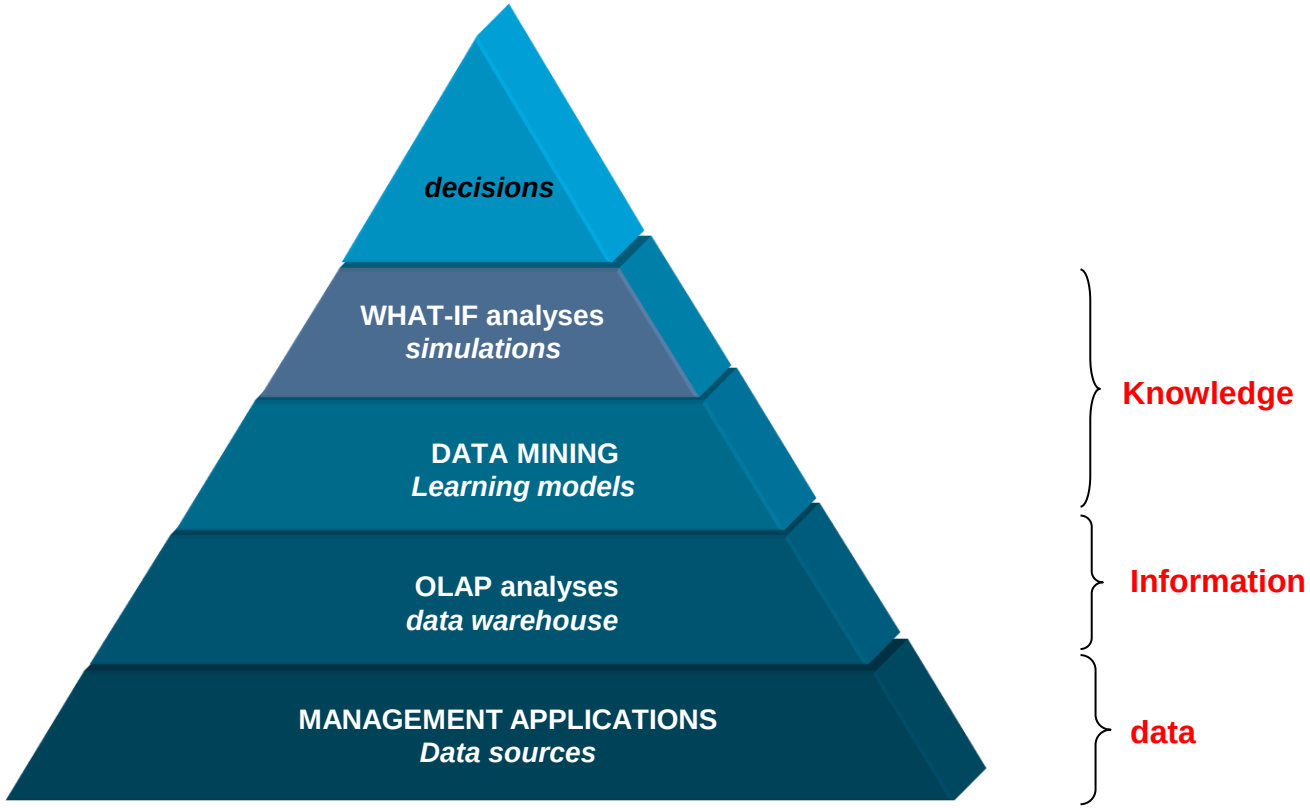
Jeremy Howard (TED talk) is an entrepreneur, business strategist, developer, and educator. He is the youngest faculty member at Singularity University, where he teaches data science. Previously he was the President and Chief Scientist of Kaggle, a community and competition platform for over 150,000 data scientists. *https://www.ted.com/talks/jeremy_howard_the_wonderful_and_terrifying_implications_of_computers_that_can_learn*

Fei Fei Li (TED talk). As Director of Stanford's Artificial Intelligence Lab and Vision Lab, Fei-Fei Li is working to solve AI's trickiest problems — including image recognition, learning and language processing. *https://www.ted.com/talks/fei_fei_li_how_we_re_teaching_computers_to_understand_pictures*

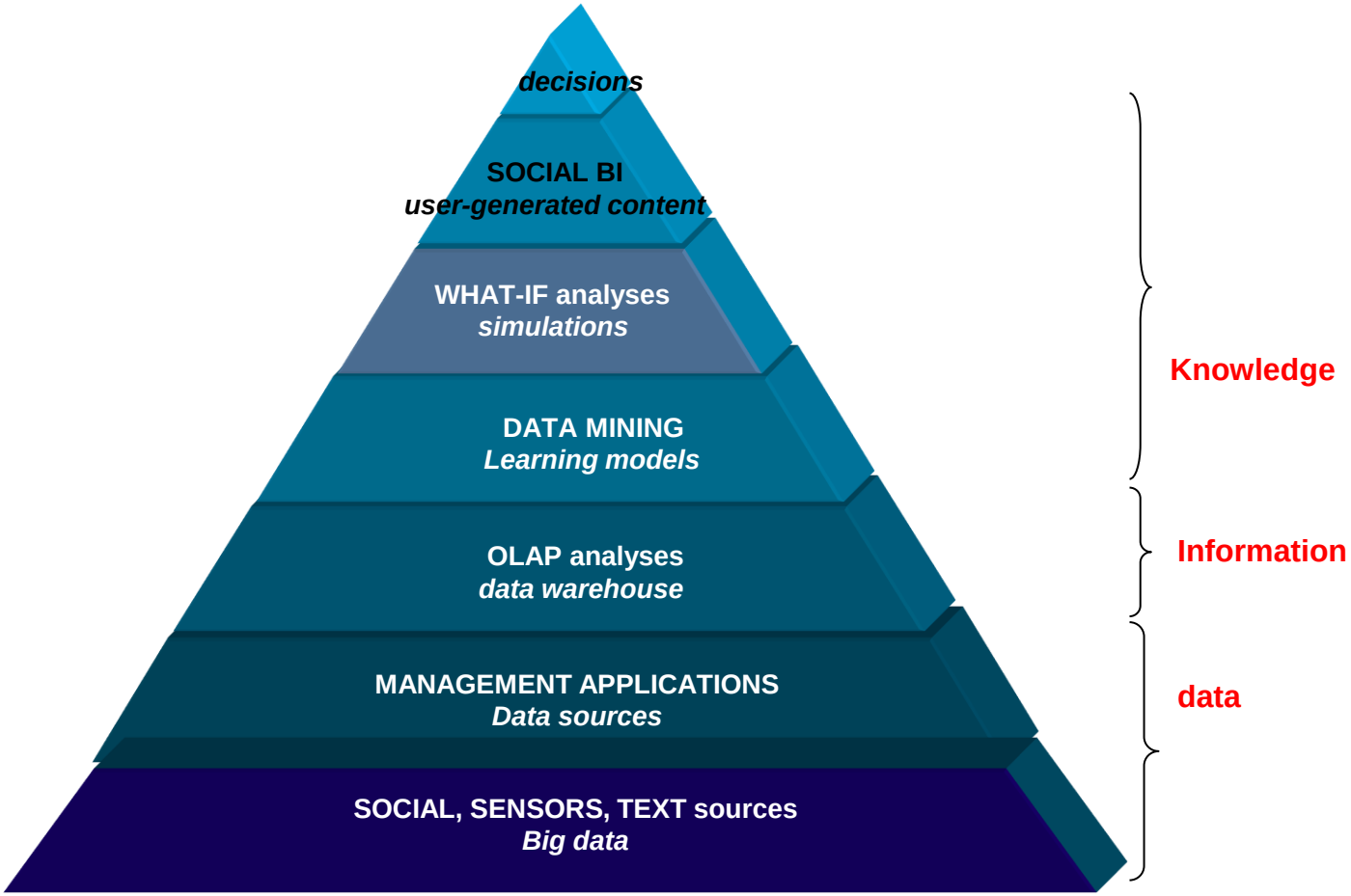
Technological Singularity

- Una singolarità tecnologica è un ipotetico punto futuro in cui il progresso tecnologico accelera oltre la nostra capacità di comprensione e previsione
- Nell'ipotesi della singolarità un agente intelligente evolvibile inizia una catena di cicli di auto-miglioramento fino a causare un'esplosione di intelligenza che crea una superintelligenza ossia un agente più intelligente dell'uomo che lo ha creato.
- Scienza o fantascienza?
  - La legge di Moore e l'evoluzione della tecnologia
  - Sono già disponibili super computer con «raw power» maggiore del cervello umano (stimato 10-100 Peta Flop).
  - La potenza di calcolo grezza non significa intelligenza!
  - Le neuro scienze sono lontane dal riuscire a riprodurre completamente il processo mentale umano

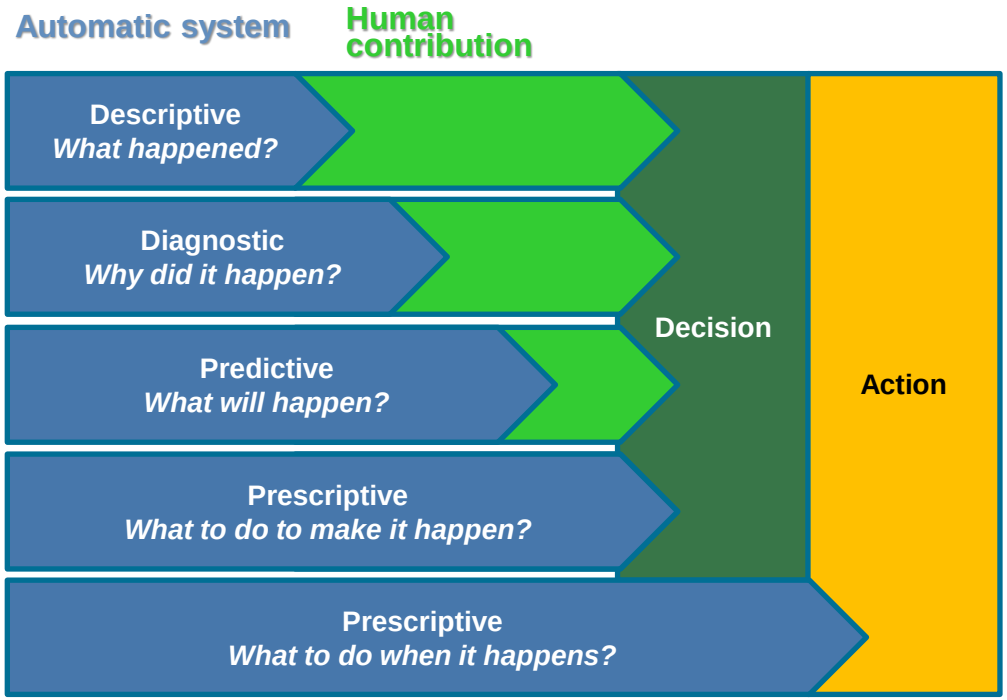
La piramide della BI 1.0



La piramide della BI 2.0

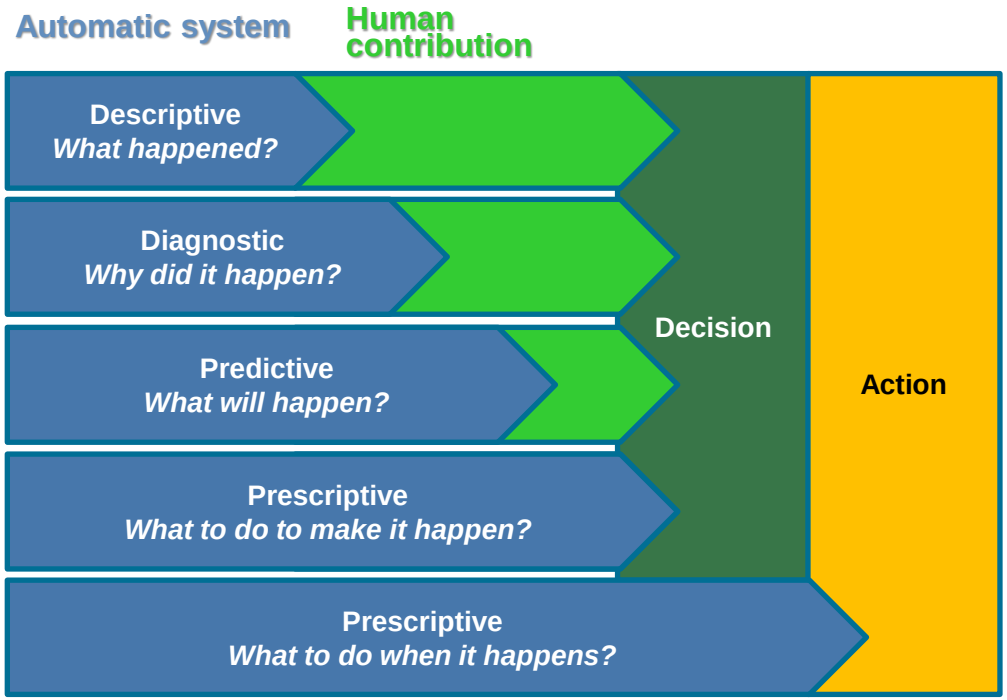


La piramide della BI 2.0



In funzione del livello di automazione di una decisione

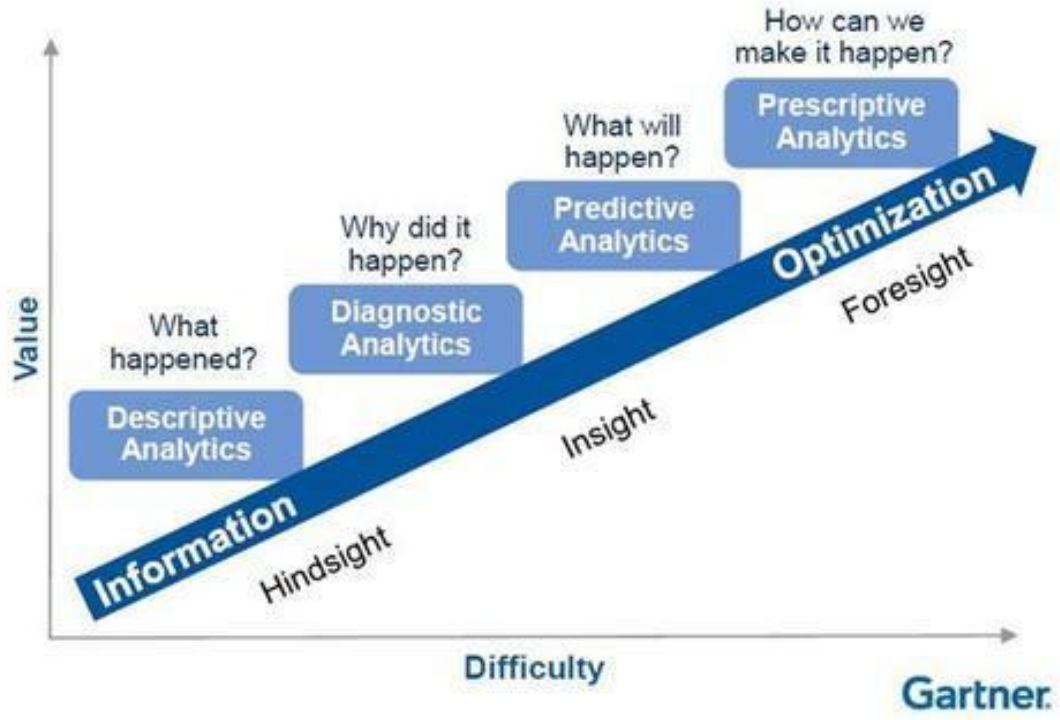
La piramide della BI 2.0



In funzione del livello di automazione di una decisione

Qual è la migliore soluzione da adottare?

La piramide della BI 2.0



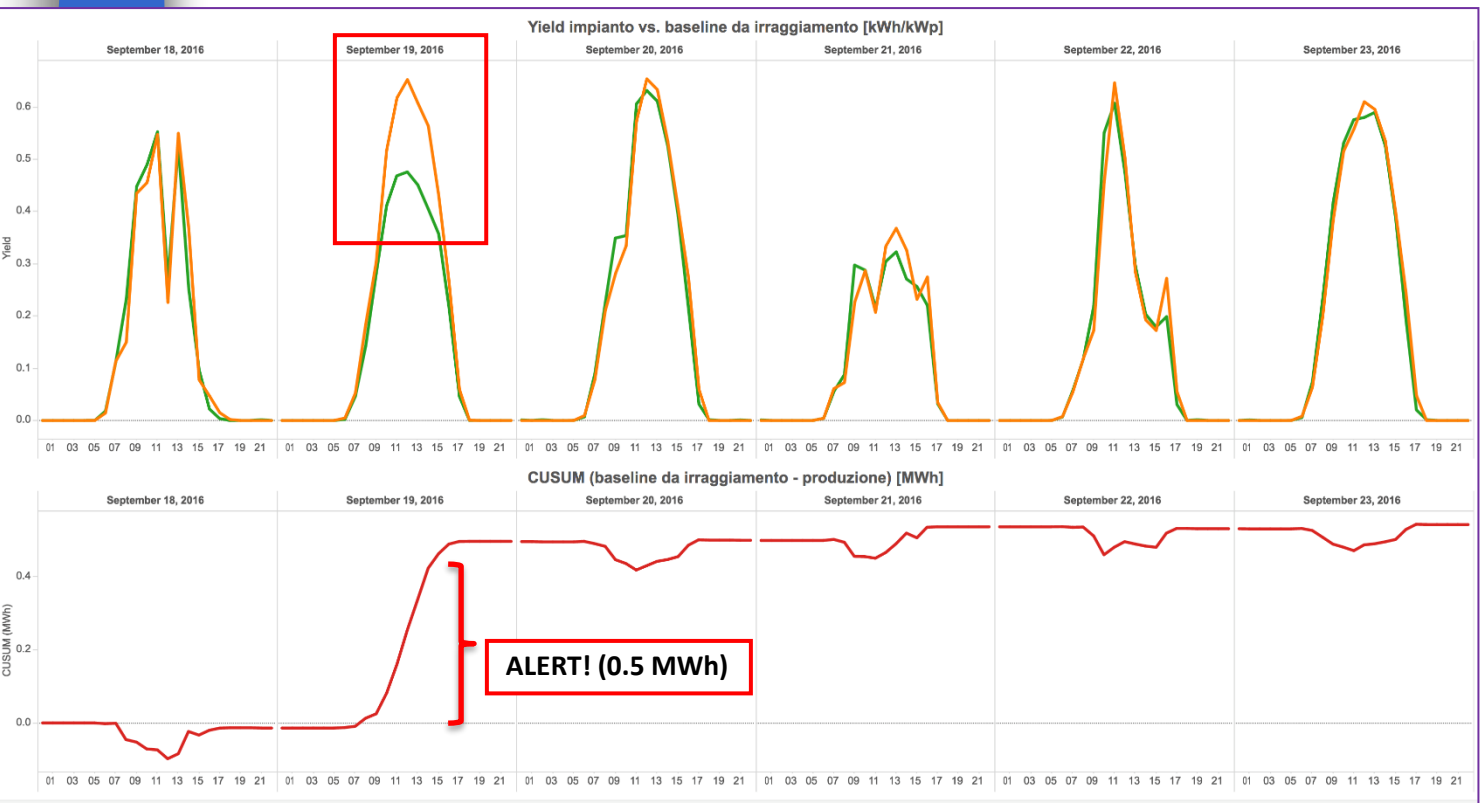
In funzione del livello di automazione di una decisione

Il più semplice che porti valore all'azienda

Descriptive analytics

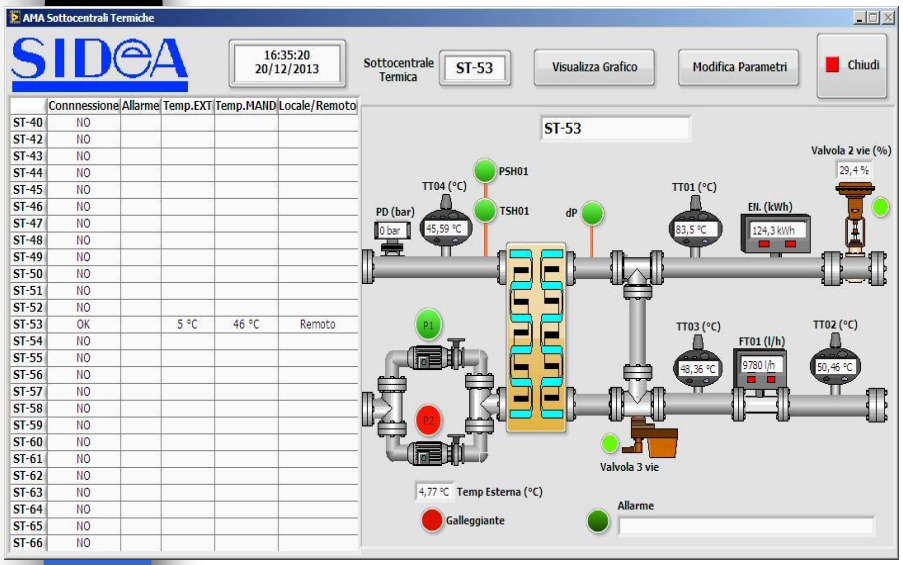
- I dati storici sono utilizzati per descrivere il sistema
- Dashboard e OLAP sono i principali tipi di visualizzazione

Energy baselines



Diagnostic analytics

- Usa i dati per capire le cause
  - ✓ Diagnosi dei guasti
  - ✓ Sistemi di allarme preventivo



Dati di impianto (dai sensori) + dati di errori:

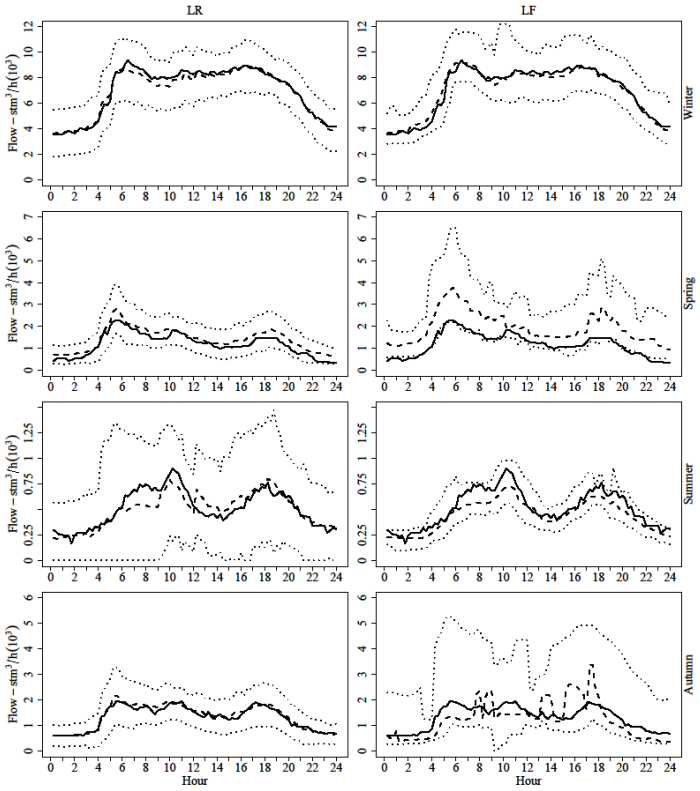
- Identifica (classifica) le serie temporali che portano agli errori
- Identificare i sensori e progettare le funzionalità che portano più informazioni (cioè sono fortemente correlate all'errore)

© 2016 National Instruments Corporation

Predictive analytics

Usa i dati per predire i comportamenti futuri

- Simulation systems
- Time series prediction
- Failure prediction
- Sales predictions



Previsione del consumo di gas per HERA

Input:

- Serie temporale di 1 anno di consumi
- Previsioni meteo giorno successivo
- Giorno della settimana

Output

- Previsione del consumo per le 24 ore successive

Prescriptive analytics

- Sistemi di ottimizzazione mono e multi goal
- Verifica scenari alternativi
- Sistemi decisionali

- Ottimizza il mix di acquisizione energetica in base alle esigenze, al mercato e alle condizioni meteorologiche
  - ✓ Il fotovoltaico è abbastanza?
  - ✓ Acquista dalla rete?
  - ✓ Biomassa?
- Decidere se irrigare dato un profilo ottimale di umidità del terreno
  - ✓ Irrigare oggi anche se piove domani?
  - ✓ Irrigare frequentemente per brevi periodi oppure irrigare abbondantemente meno frequentemente

Il percorso di adozione della BI

Il percorso di adozione della BI è incrementale e raramente permette di saltare dei passaggi
E' *rischioso*, *costoso* e *inutile* adottare soluzioni avanzate senza avere completamente sfruttato quelle semplici

- I manager non sono pronti
  - ✓ Non hanno il mindset giusto
- I dati non sono pronti
  - ✓ Non sono di qualità sufficiente
- I processi delle aziende non sono pronte
  - ✓ Non sono definiti in modo da appoggiarsi ai dati e di reagire a essi

Dubitate dei consulenti e dei fornitori di software che offrono analisi avanzate se la vostra azienda sfrutta a malapena il data warehouse aziendale

Creare aziende data-driven

Il termine *aziende data-driven* si riferisce ad aziende in cui le decisioni e i processi sono supportate dai dati

- Le decisioni si basano su conoscenze quantitative piuttosto che qualitative
- I processi e le conoscenze sono una risorsa dell'azienda e non vanno persi se i manager cambiano

La differenza tra una decisione basata sui dati e una buona decisione è un buon manager

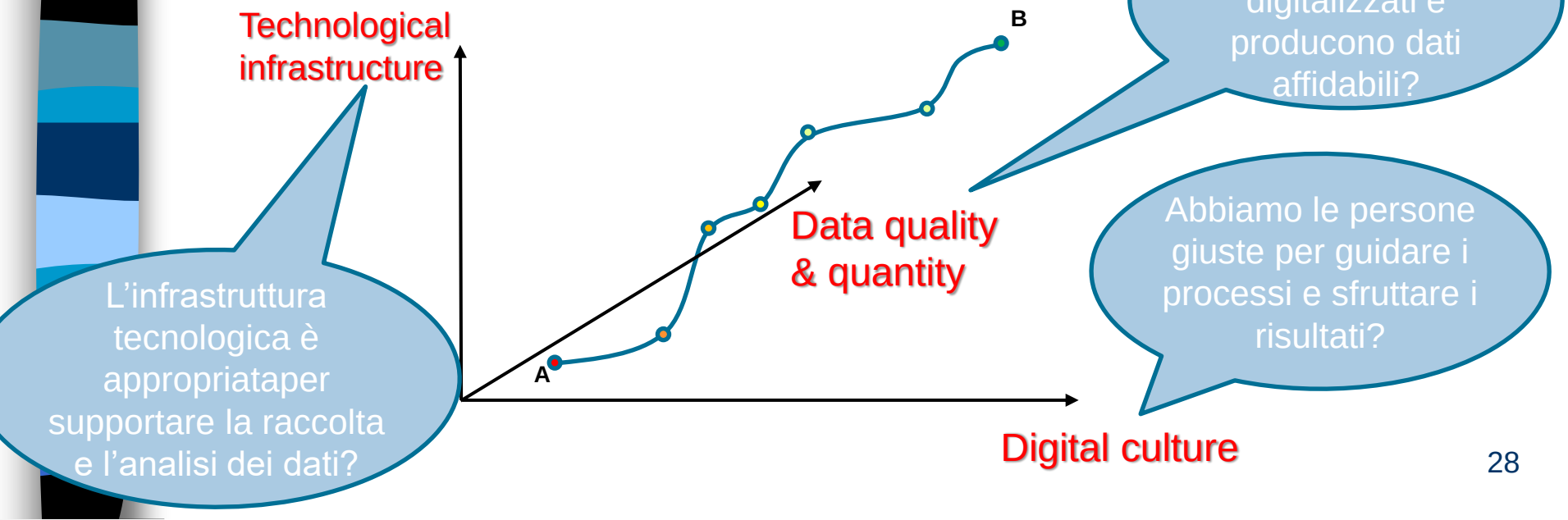
L'adozione di una mentalità basata sui dati va ben oltre l'adozione di una soluzione di business intelligence e comporta:

- ✓ **Creare una cultura dei dati**
- ✓ **Cambiare la mentalità dei manager**
- ✓ **Cambiare i processi**
- ✓ **Migliorare la qualità di tutti i dati**

Creare aziende data-driven

Quello di digitalizzazione è un percorso che coinvolge tre dimensioni principali. Passare da A a B è un processo pluriennale fatto di obiettivi intermedi, ciascuno dei quali deve essere raggiungibile

- Deve risolvere un problema e apportare valore
- Deve essere realizzabile in un intervallo di tempo limitato (in genere meno di un anno)
- I costi devono essere economicamente correlati agli utili
- …..



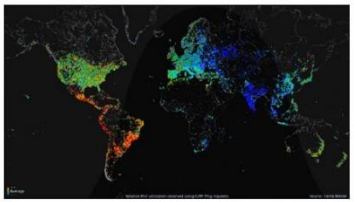
La Data Revolution

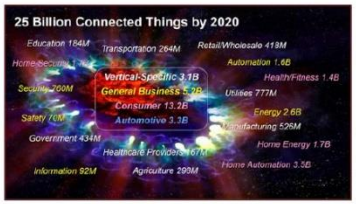
- I dati rappresentano il principale combustibile che alimenta la trasformazione digitale
- La digitalizzazione è iniziata negli anni ‘70s con la progressiva diffusione dei calcolatori dando il via al processo di digitalizzazione dei processi e delle Informazioni che continua ad accelerare ancora oggi cambiando nome ma non obiettivo
  - Post-industrial society
  - Information technology revolution
  - Digital age

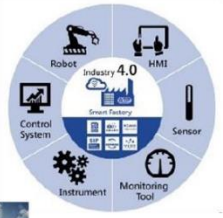
- Possiamo stimare l'inizio della **Digital Age** nel 2002, quando nel mondo sono state archiviate più informazioni digitali che analogiche.
  - Alla fine degli anni '80 meno dell'1% delle informazioni era in formato digitale,
  - Nel 2012 la percentuale era salita al 99% con un incremento annuo di circa il 30%, che porta ad un raddoppio delle informazioni conservate in meno di 3 anni.

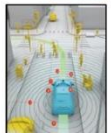
Chi produce i dati nella digital age?

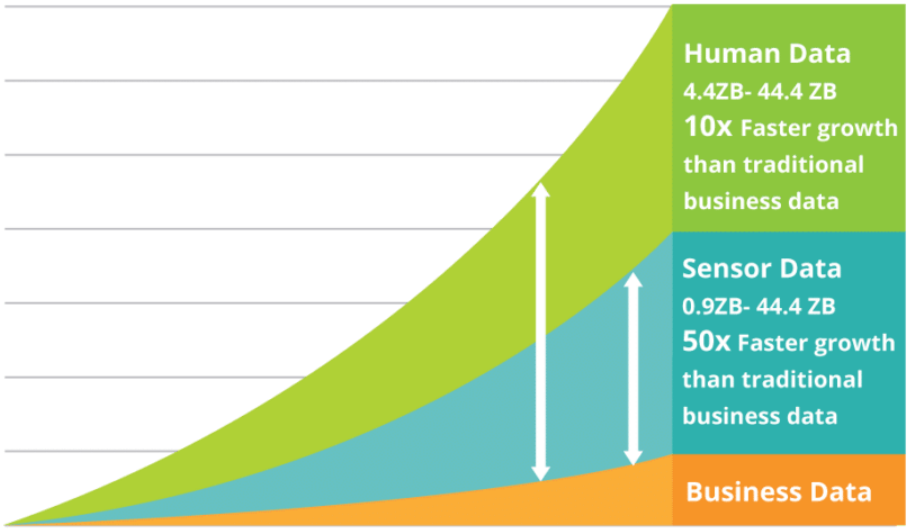
- I sistemi informativi non sono più limitati ai dati prodotti dai processi aziendali ma vanno ripensati per permettere di sfruttare tutti i dati utili all'azienda e per poter supportare processi interni ed esterni





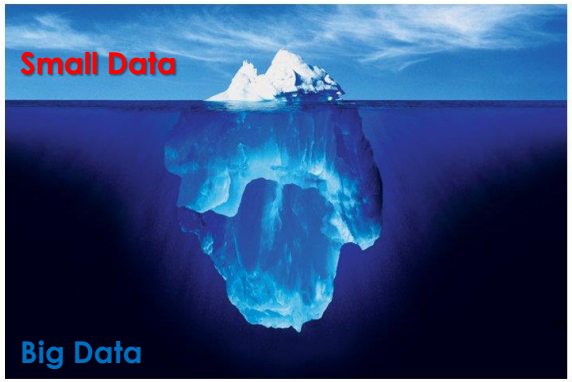






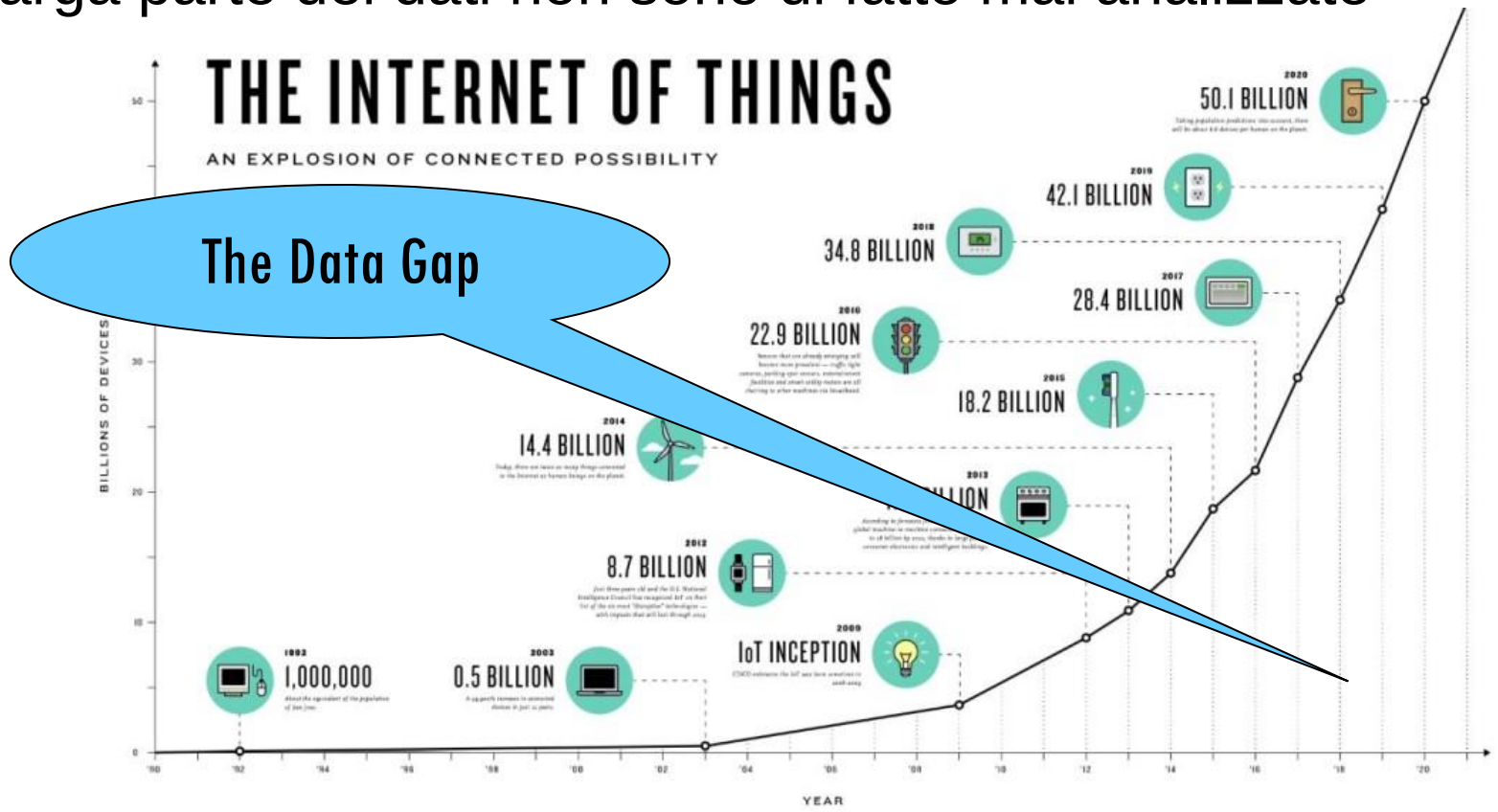
Big Data vs Small Data

- La progressiva digitalizzazione di servizi e impianti genera una enorme massa di dati eterogenei e in tempo reale
- I Big Data devono essere trasformati in Small data affinché possano essere sfruttati ai fini decisionali
- Per gestire questa trasformazione occorrono
  - ✓ Tecnologia ad hoc (NO SQL DBMS)
  - ✓ Potenza di calcolo (cluster computing)
  - ✓ Sistemi automatizzati (Intelligenza artificiale)



Data mining su grandi data set

- Molte delle informazioni presenti sui dati non sono direttamente evidenti
- Le analisi guidate dagli uomini possono richiedere settimane per scoprire informazioni utili
- Larga parte dei dati non sono di fatto mai analizzate

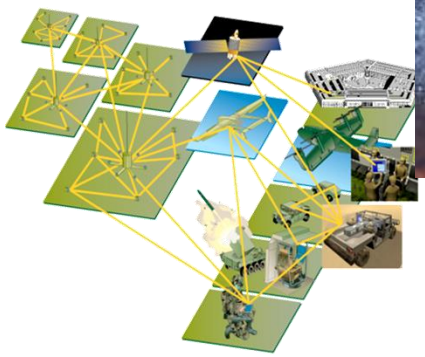


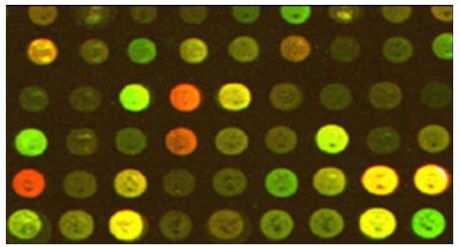
Perché fare data mining?

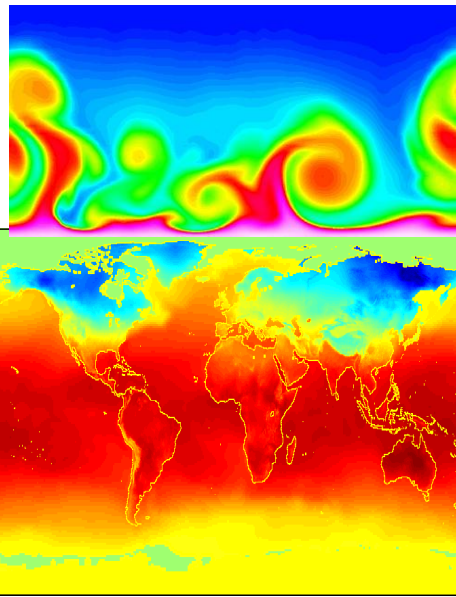
- La quantità dei dati memorizzata su supporti informatici è in continuo aumento
  - ✓ Pagine Web, sistemi di e-commerce
  - ✓ Dati relativi ad acquisti/scontrini fiscali
  - ✓ Transazioni bancarie e relative a carte di credito
- L'hardware diventa ogni giorno più potente e meno costoso
- La pressione competitiva è in continua crescita
  - ✓ La risorsa informazione è un bene prezioso per superare la concorrenza

Perché fare data mining?

- I dati prodotti e memorizzati crescono a grande velocità (GB/ora)
  - ✓ Sensori posti sui satelliti
  - ✓ Telescopi
  - ✓ Microarray che generano espressioni genetiche
  - ✓ Simulazioni scientifiche che producono terabyte di dati
- Le tecniche tradizionali sono inapplicabili alle masse di dati grezzi
- Il Data mining può aiutare gli scienziati a:
  - ✓ Classificare e segmentare i dati
  - ✓ Formulare ipotesi

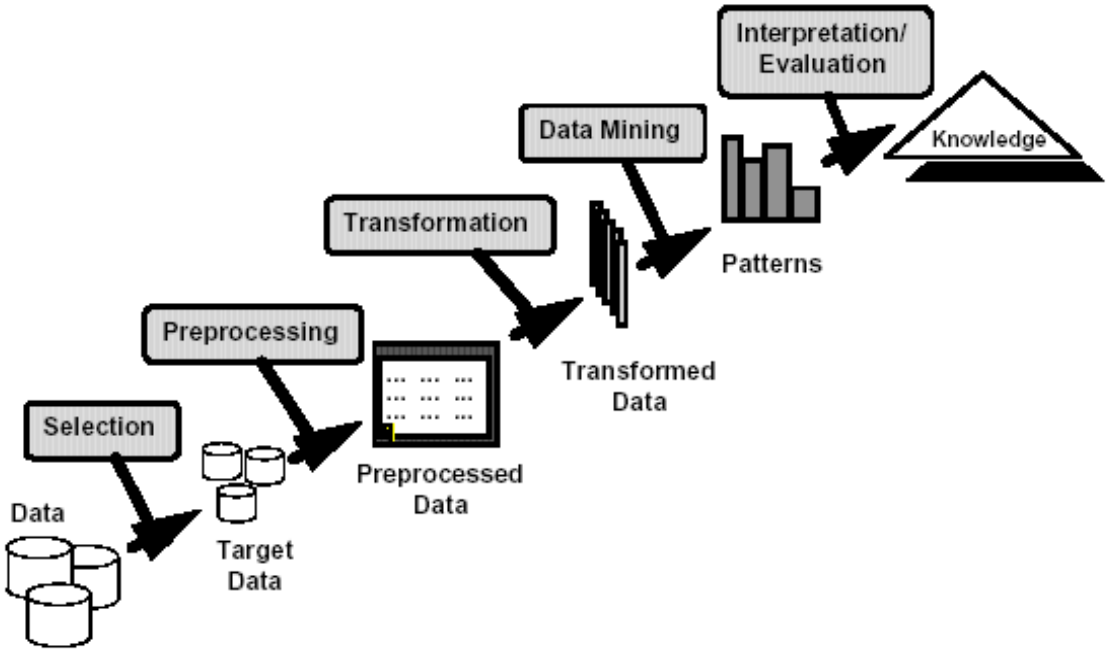






Cosa è il Data Mining?

- Alcune definizioni
  - ✓ Estrazione complessa di informazioni implicite, precedentemente sconosciute e potenzialmente utili dai dati.
  - ✓ Esplorazione e analisi, per mezzo di sistemi automatici e semi-automatici, di grandi quantità di dati al fine di scoprire pattern significativi

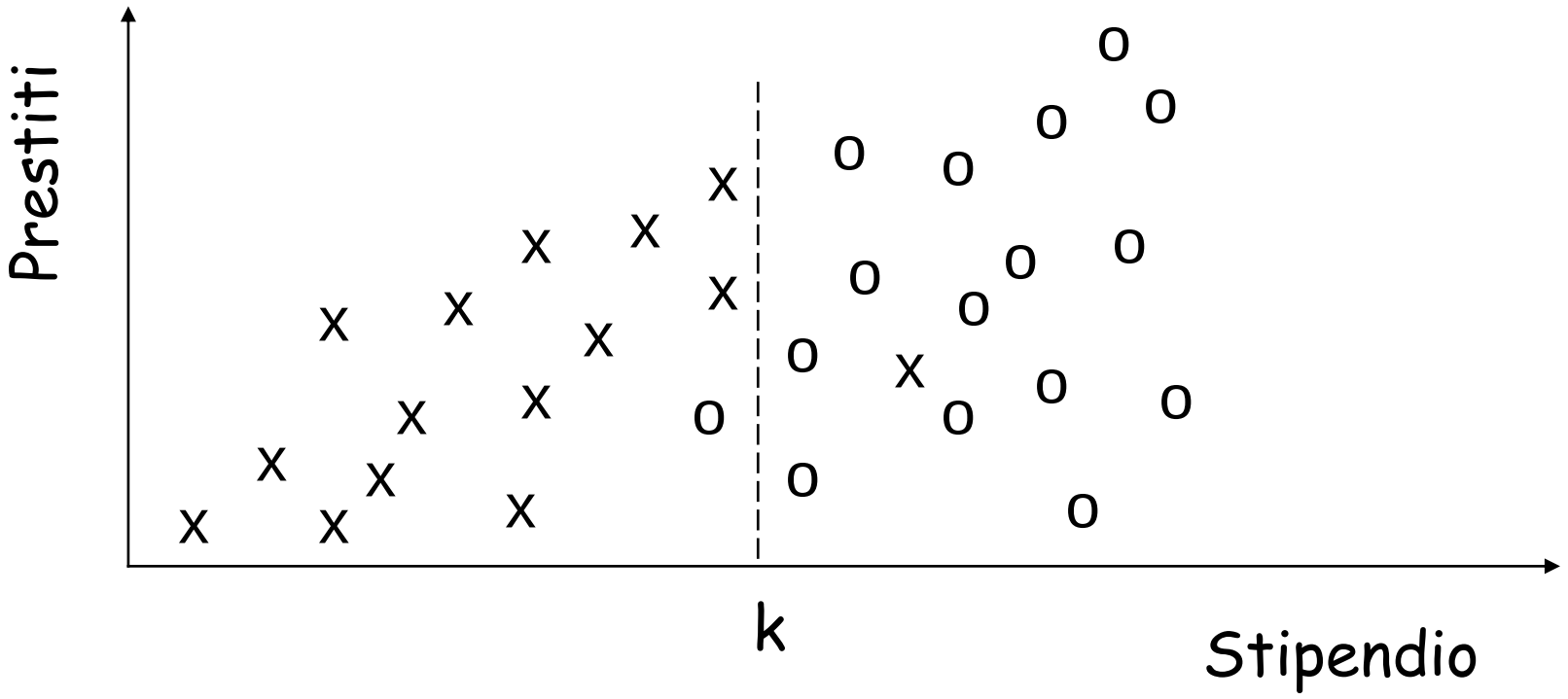


Pattern

- Un pattern è una rappresentazione sintetica e ricca di semantica di un insieme di dati; esprime in genere un modello ricorrente nei dati, ma può anche esprimere un modello eccezionale
- Un pattern deve essere:
  - ✓ Valido sui dati con un certo grado di confidenza
  - ✓ Comprensibile dal punto di vista sintattico e semantico, affinché l'utente lo possa interpretare
  - ✓ Precedentemente sconosciuto e potenzialmente utile, affinché l'utente possa intraprendere azioni di conseguenza

Esempio

Persone che hanno ricevuto un prestito
- x: hanno mancato la restituzione di rate
- o: hanno rispettato le scadenze



- Pattern:
  - ✓ **IF** stipendio < k **THEN** pagamenti mancati

Tipi di pattern

- Regole associative
  - ✓ consentono di determinare le regole di implicazione logica presenti nella base di dati, quindi di individuare i gruppi di affinità tra oggetti
- Classificatori
  - ✓ consentono di derivare un modello per la classificazione di dati secondo un insieme di classi assegnate a priori
- Alberi decisionali
  - ✓ sono particolari classificatori che permettono di identificare, in ordine di importanza, le cause che portano al verificarsi di un evento
- Clustering
  - ✓ raggruppa gli elementi di un insieme, a seconda delle loro caratteristiche, in classi non assegnate a priori
- Serie temporali
  - ✓ Permettono l'individuazione di pattern ricorrenti o atipici in sequenze di dati complesse

Cosa NON è Data Mining?

Cosa NON è Data Mining?

– Cercare un numero nell'elenco telefonico

– Interrogare un motore di ricerca per cercare informazioni su "Amazon"

Cosa è Data Mining?

–Certi cognomi sono più comuni in certe regioni (es. Casadei, Casadio, … in Romagna)

– Raggruppare i documenti restituiti da un motore di ricerca in base a informazioni di contesto (es. "Amazon rainforest", "Amazon.com")

Cosa NON è Data Mining?

SQL is sufficient

Cosa NON è Data Mining?

– Cercare un numero nell'elenco telefonico

– Interrogare un motore di ricerca per cercare informazioni su "Amazon"

We look for correlation between surnames and all the person attributes

Cosa è Data Mining?

–Certi cognomi sono più comuni in certe regioni (es. Casadei, Casadio, … in Romagna)

– Raggruppare i documenti restituiti da un motore di ricerca in base a informazioni di contesto (es. "Amazon rainforest", "Amazon.com")

Cosa NON è Data Mining?

SQL is sufficient

Cosa NON è Data Mining?

– Cercare un numero nell'elenco telefonico

– Interrogare un motore di ricerca per cercare informazioni su "Amazon"

It is a typical information retrieval task

We look for correlation between surnames and all the person attributes

Cosa è Data Mining?

–Certi cognomi sono più comuni in certe regioni (es. Casadei, Casadio, … in Romagna)

– Raggruppare i documenti restituiti da un motore di ricerca in base a informazioni di contesto (es. "Amazon rainforest", "Amazon.com")

Requires a semantic understanding of the text through NPL and onthologies

Le origini del Data Mining

- Questa disciplina trae ispirazioni dalle aree del machine learning/intelligenza artificiale, pattern recognition, statistica e basi di dati
- Le tradizionali tecniche di analisi risultano inidonee per molteplici motivi
  - ✓ Quantità dei dati
  - ✓ Elevata dimensionalità dei dati
  - ✓ Eterogeneità dei dati

Attività tipiche del Data Mining

- Sistemi di predizione
  - ✓ Utilizzare alcune variabili per predire il valore incognito o futuro di altre variabili.
- Sistemi di descrizione
  - ✓ Trovare pattern interpretabili dall'uomo che descrivano I dati

Attività tipiche del Data Mining

- Classificazione [Predittiva]
- Clustering [Descrittiva]
- Ricerca di regole associative [Descrittiva]
- Ricerca di pattern sequenziali [Descrittiva]
- Regressione [Predittiva]
- Individuazione di deviazioni [Predittiva]

Machine learning

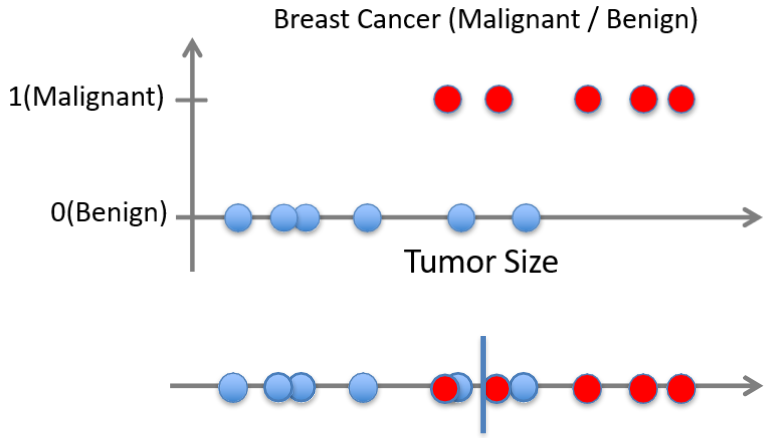
- Supervised (inductive) learning
  - ✓ Given: training data + desired outputs (labels)
- Unsupervised learning
  - ✓ Given: training data (without desired outputs)
- Semi-supervised learning
  - ✓ Given: training data + a few desired outputs
- Reinforcement learning
  - ✓ Rewards from sequence of actions

Machine learning

- Supervised learning tasks
  - ✓ The training set you feed to the algorithm includes the desired solutions, called labels
  - ✓ Classification
    - Approximating a mapping function (f) from input variables (X) to discrete output variables (y)
    - The output variables are called labels or categories
    - The mapping function predicts the class or category for a given observation
    - E.g., a spam filter is trained with many example emails along with their class (spam or ham)
  - ✓ Regression
    - Approximating a mapping function (f) from input variables (X) to a continuous output variable (y)
    - A continuous output variable is a real-value, such as an integer or floating-point value
    - E.g., predict the price of a car given a set of features (mileage, age, brand, etc.) called predictors

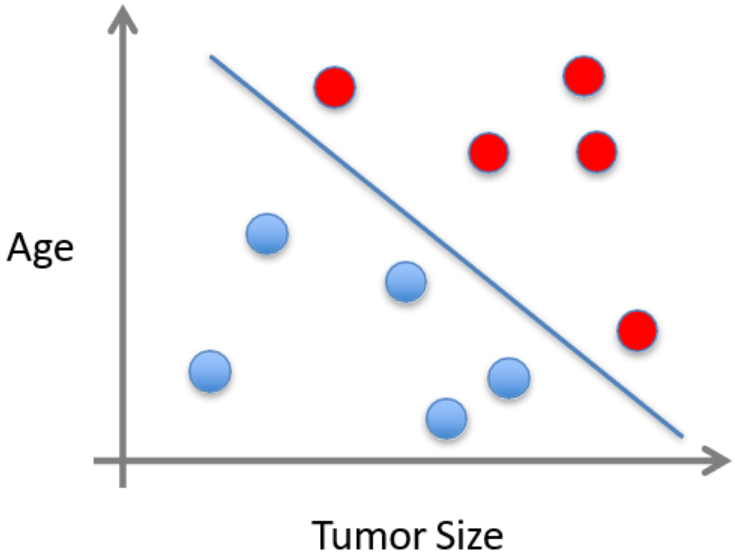
Supervised Learning: Classification

- Classification
  - ✓ Given $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
  - ✓ Learn a function f(x) to predict y given x where y is categorical



Supervised Learning: Classification

- x can be multi-dimensional
  - ✓ Each dimension corresponds to an attribute/column of the dataset



Classificazione: Definizione

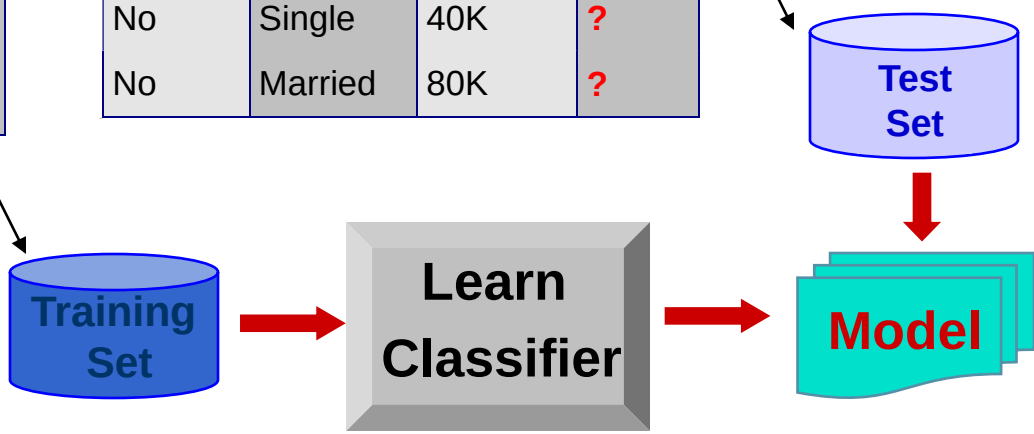
- Data una collezione di record (*training set*)
  - ✓ Ogni record è composto da un insieme di *attributi*, di cui uno esprime la *classe* di appartenenza del record.
- Trova un *modello* per l'attributo di classe che esprima il valore dell'attributo in funzione dei valori degli altri attributi.
- Obiettivo: record non noti devono essere assegnati a una classe nel modo più accurato possibile
  - ✓ Viene utilizzato un *test set* per determinare l'accuratezza del modello. Normalmente, il data set fornito è suddiviso in training set e test set. Il primo è utilizzato per costruire il modello, il secondo per validarlo.

Classificazione: Esempio

categorical categorical continuous class

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Classificazione: Applicazione 1

- Direct Marketing
  - ✓ Obiettivo: Ridurre il costo della pubblicità via posta *definendo* l'insieme dei clienti che, con maggiore probabilità, compreranno un nuovo prodotto di telefonia
  - ✓ Approccio:
    - Utilizza i dati raccolti per il lancio di prodotti similari
    - Conosciamo quali clienti hanno deciso di comprare e quali no Questa informazione *{compra, non compra}* rappresenta *l'attributo di classificazione*
    - Raccogli tutte le informazioni possibili legate ai singoli compratori: demografiche, stile di vita, precedenti rapporti con l'azienda
      - Attività lavorativa svolta, reddito, età, sesso, ecc.
    - Utilizza queste informazioni come attributi di input per addestrare un modello di classificazione

Classificazione: Applicazione 2

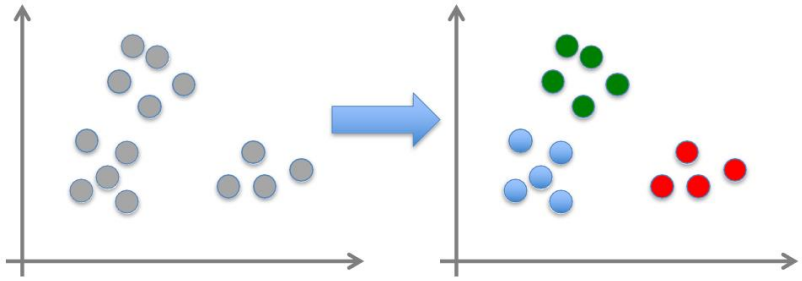
- Individuazione di frodi
  - ✓ Obiettivo: predire l'utilizzo fraudolento delle carte di credito
  - ✓ Approccio:
    - Utilizza le precedenti transazioni e le informazioni sui loro possessori come attributi
      - Quando compra l'utente, cosa compra, paga con ritardo, ecc.
    - Etichetta le precedenti transazioni come fraudolenti o lecite
    - Questa informazione rappresenta l'attributo di classificazione
    - Costruisci un modello per le due classi di transazioni
    - Utilizza il modello per individuare comportamenti fraudolenti delle prossime transazioni relative a una specifica carta di credito

Classificazione: Applicazione 3

- Individuazione dell'insoddisfazione del cliente:
  - ✓ Obiettivo: Predire clienti propensi a passare a un concorrente.
  - ✓ Approccio:
    - Utilizza i dati relativi agli acquisti dei singoli utenti (presenti e passati) per trovare gli attributi rilevanti
      - Quanto spesso l'utente contatta l'azienda, dove chiama, in quali ore del giorno chiama più di frequente, quale è la sua situazione finanziaria, è sposato, ecc.
    - Etichetta gli utenti come fedeli o non fedeli
    - Trova un modello che definisca la fedeltà

Unsupervised Learning

- ✓ Given x_1, x_2, ..., x_n (without labels)
- ✓ Output hidden structure behind the x's
  - E.g., clustering



Clustering: Definizione

- Dato un insieme di punti, ognuno caratterizzato da un insieme di attributi, e avendo a disposizione una *misura di similarità* tra i punti, trovare i sottoinsiemi di punti tali che:
  - ✓ I punti appartenenti a un sottoinsieme sono più simili tra loro rispetto a quelli appartenenti ad altri cluster

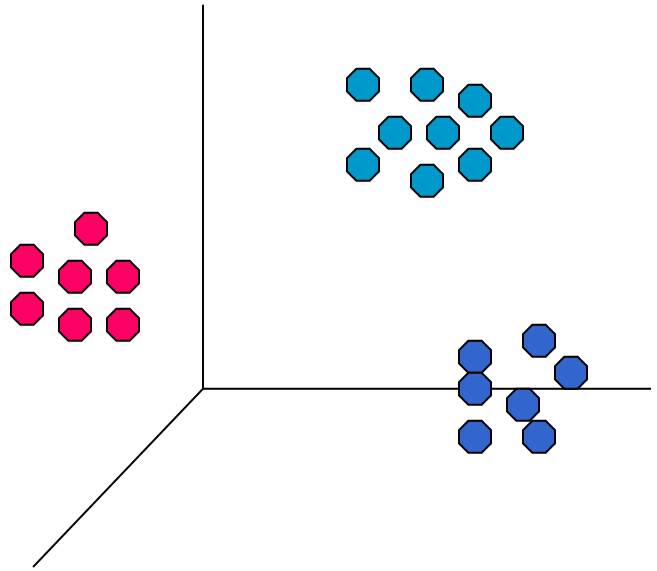
  - Misure di similarità:
  - ✓ La distanza euclidea è applicabile se gli attributi dei punti assumono valori continui
  - ✓ Sono possibili molte altre misure che dipendono dal problema in esame

Rappresentazione del clustering

- Rappresentazione di un clustering nello spazio 3d costruito utilizzando la distanza euclidea come misura di similarità

Le distanze intra-cluster sono minimizzate

Le distanze inter-cluster sono massimizzate



Clustering: Applicazione 1

- Segmentazione del mercato:
  - ✓ Obiettivo: suddividere i clienti in sottoinsiemi distinti da utilizzare come target di specifiche attività di marketing
  - ✓ Approccio:
    - Raccogliere informazioni sui clienti legati allo stile di vita e alla collocazione geografica
    - Trovare cluster di clienti simili
    - Misurare la qualità dei cluster verificando se il pattern di acquisto dei clienti appartenenti allo stesso cluster è più simile di quello di clienti appartenenti a cluster distinti

Clustering: Applicazione 2

- Clustering di documenti:
  - ✓ Obiettivo: trovare sottogruppi di documenti che sono simili sulla base dei termini più rilevanti che in essi compaiono
  - ✓ Approccio: Identificare i termini che si presentano con maggiore frequenza nei diversi documenti. Definire una misura di similarità basata sulla frequenza dei termini e usarla per creare i cluster.

Clustering di documenti

- Punti da clusterizzare: 3204 articoli del Los Angeles Times.
- Misura di similarità: numero di parole comuni tra due documenti (escluse alcune parole comuni).

Categoria	*# articoli*	*#correttamente classsificati*	*%correttamente classsificati*
Finanza	555	364	66%
Esteri	341	260	76%
Cronaca nazionale	273	36	13%
Cronaca locale	943	746	79%
Sport	738	573	78%
Intrattenimento	354	278	79%

Regole associative: Definizione

- Dato un insieme di record ognuno composto da più elementi appartenenti a una collezione data
  - ✓ Produce delle regole di dipendenza che predicono l'occorrenza di uno degli elementi in presenza di occorrenze degli altri.

TID	Record
1	Pane, Coca Cola, Latte
2	Birra, Pane
3	Birra, Coca Cola, Pannolini, Latte
4	Birra, Pane, Pannolini, Latte
5	Birra, Pannolini, Latte

Regola:

{Latte} --> {Coca Cola}

{Pannolini, Latte} --> {Birra}

Regole associative: applicazione 1

- Marketing e promozione delle vendite:
  - ✓ Si supponga di avere scoperto la regola associativa

    {Bagels, … } --> {Potato Chips}
  - ✓ Potato Chips come conseguente: l'informazione può essere utilizzata per capire quali azioni intraprendere per incrementare le sue vendite
  - ✓ Bagels come antecedente: l'informazione può essere utilizzata per capire quali prodotti potrebbero essere condizionati nel caso in cui il negozio interrompesse la vendita dei Bagel

Regole associative: Applicazione 2

- Disposizione della merce.
  - ✓ Obiettivo: identificare i prodotti comprati assieme da un numero sufficientemente elevato di clienti.
  - ✓ Approccio: utilizza i dati provenienti dagli scontrini fiscali per individuare le dipendenze tra i prodotti.
  - ✓ Una classica regola associativa
    - Se un cliente compra pannolini e latte, allora molto probabilmente comprerà birra.
    - Quindi non vi stupite se trovate le casse di birra accanto ai pannolini!

Regole associative: Applicazione 3

- Gestione dell'inventario:
  - ✓ Obiettivo: un'azienda che effettua riparazione di elettrodomestici vuole studiare le relazioni tra i malfunzionamenti denunciati e i ricambi richiesti al fine di equipaggiare correttamente i propri veicoli e ridurre le visite alle abitazioni dei clienti.
  - ✓ Approccio: elabora i dati relativi ai ricambi utilizzati nei precedenti interventi alla ricerca di pattern di co-occorrenza.

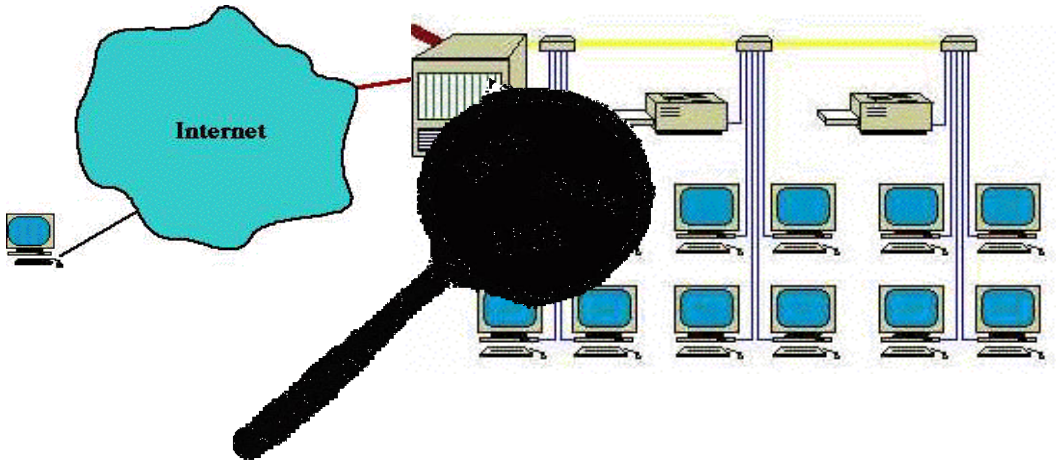
Regressione

- Predire il valore di una variabile a valori continui sulla base di valori di altre variabili assumendo un modello di dipendenza lineare/non lineare.
- Problema ampiamente studiato in statistica e nell'ambito delle reti neurali.
- Esempi:
  - ✓ Predire il fatturato di vendita di un nuovo prodotto sulla base degli investimenti in pubblicità.
  - ✓ Predire la velocità del vento in funzione della temperatura, umidità, pressione atmosferica
  - ✓ Predire l'andamento del mercato azionario.

Identificazione di comportamenti anomali e scostamenti

- Identificazione di scostamenti dal normale comportamento
- Applicazioni:
  - ✓ Identificazioni di frodi nell'uso delle carte di credito

  - ✓ Identificazioni di intrusioni in rete

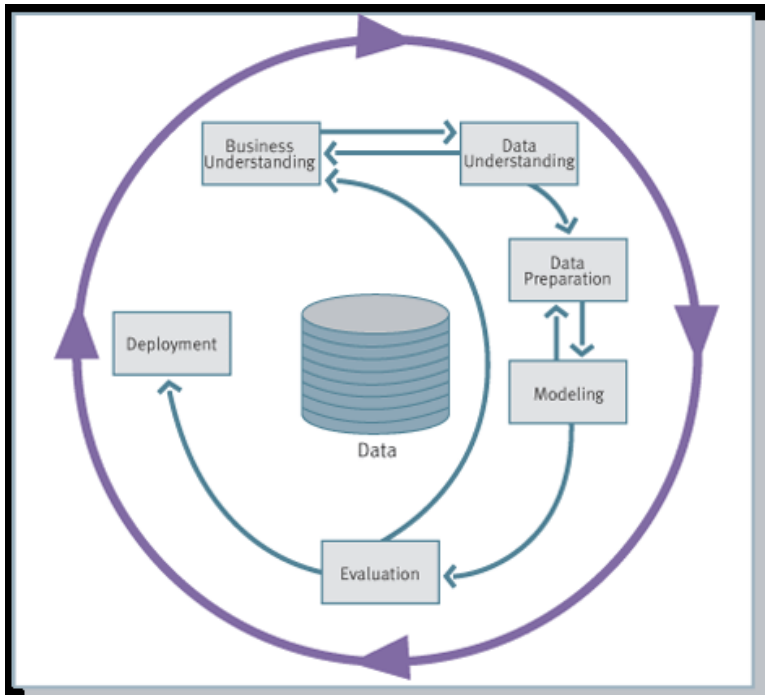


Scommesse del Data Mining

- Scalabilità
- Multidimensionalità del data set
- Complessità ed eterogeneità dei dati
- Qualità dei dati
- Proprietà dei dati
- Mantenimento della privacy
- Processing in real time

CRISP-DM: un approccio metodologico

- Un progetto di Data mining richiede un approccio strutturato in cui la scelta del miglior algoritmo è solo uno dei fattori di successo
- La metodologia **CRISP-DM** (*Cross-Industry Standard Process for Data Mining*) è una delle proposte maggiormente strutturate per definire i passi fondamentali di un progetto di Data Mining
- Le sei fasi del ciclo di vita non sono strettamente sequenziali. Tornare su attività già svolte è spesso necessario



CRISP-DM: le fasi

1) Comprensione del dominio applicativo: capire gli obiettivi del progetto dal punto di vista dell'utente, tradurre il problema dell'utente in un problema di data mining e definire un primo piano di progetto

2) Comprensione dei dati: raccolta preliminare dei dati finalizzata a identificare problemi di qualità e a svolgere analisi preliminari che permettano di identificarne le caratteristiche salienti

3) Preparazione dei dati: comprende tutte le attività necessarie a creare il dataset finale: selezione di attributi e record, trasformazione e pulizia dei dati

CRISP-DM: le fasi

4) Creazione del modello: diverse tecniche di data mining sono applicate al dataset anche con parametri diversi al fine di individuare quella che permette di costruire il modello più accurato

4) Valutazione del modello e dei risultati: il modello/i ottenuti dalla fase precedente sono analizzati al fine di verificare che siano sufficientemente precisi e robusti da rispondere adeguatamente agli obiettivi dell'utente

5) Deployment: il modello costruito e la conoscenza acquisita devono essere messi a disposizione degli utenti. Questa fase può quindi semplicemente comportare la creazione di un report oppure può richiedere di implementare un sistema di data mining controllabile direttamente dall'utente