

SUPSI

Data Science

Lezione 1 Introduzione

Alessandro Giusti – Dario Piga

IDSIA – SUPSI, Galleria 1, Manno

alessandro.giusti@supsi.ch

dario.piga@supsi.ch

Parte del materiale tratto da:

F. Stella, “Business Intelligence”, corso di Laurea in Informatica, Università degli studi di Milano - Bicocca

L. Azzimonti. Slide per il corso di “Business Intelligence”. Corso di Laurea in Ing. Gestionale. SUPSI

Programma del corso

Prima parte

- Introduzione alla Data Science
- Analisi esplorativa dei dati: statistiche univariate e bivariate
- Tecniche di visualizzazione

Seconda parte

- Supervised learning
- Unsupervised learning

Laboratori e strumenti software



Valutazione

- 1/3 della nota: primo compitino, inizio novembre (da definire)
- 1/3 della nota: secondo compitino, fine corso
- 1/3 lavoro di progetto su un argomento a scelta (da concordare)

Materiale Didattico

- Slide
- Link proposti di volta in volta su iCorsi
- Codice mostrato in aula

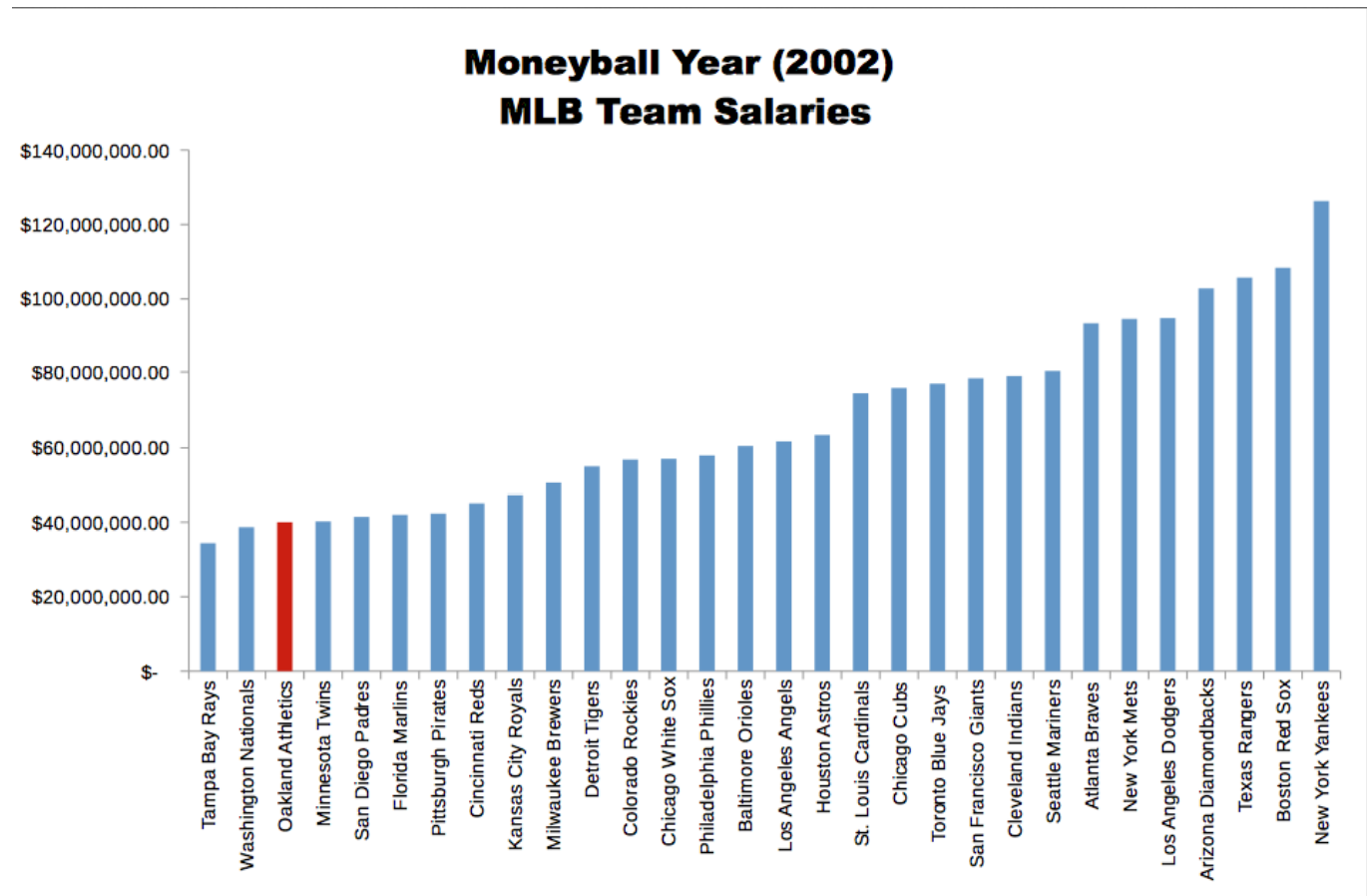
Tutto il materiale sara' disponibile su iCorsi (Data Science) al piu' tardi il giorno dopo ogni lezione. Enrolment key: **ds2018**

Link al corso: **<https://bit.ly/2NQCjuW>**

Portare il laptop a lezione, sono previste attivita' di laboratorio

Moneyball: The Art of Winning an Unfair Game (2011)

Ovvero come vincere il campionato di baseball se non hai soldi per pagare giocatori famosi



Sabermetrica

Approccio analitico, evidence-based alle decisioni di management

Il trailer di Moneyball (L'arte di vincere), 2011

<https://www.youtube.com/watch?v=IT9fyMoh6lY>

La scena con i nerd

<https://www.youtube.com/watch?v=KWPhV6PUr9o>



Billy Beane,
manager degli
Oakland Athletics
nel 2001

Sabermetrica

Dal Manifesto della sabermetrica di David Grabiner:

« Bill James ha definito la sabermetrica come “la ricerca per la conoscenza oggettiva sul baseball.” Così la sabermetrica cerca di rispondere alle domande oggettive sul baseball come “quale giocatore dei Boston Red Sox contribuisce di più all'attacco nella sua squadra?” o “quanti fuoricampi farà Ken Griffey, Jr. il prossimo anno?”. Non può trattare delle domande soggettive che sono comunque importanti per il gioco, come “qual è il tuo giocatore preferito?” »

Jackie Robinson																					
Jackie Robinson Hitting Stats																					
Yr	Age	Team	G	AB	R	H	2B	3B	HR	GRSL	RBI	BB	IBB	SO	SH	SF	HBP	GIDP	AVG	OBP	SLG
1947	28	Dodgers	151	590	125	175	31	5	12	0	48	74	0	36	28	-	9	5	.297	.383	.427
1948	29	Dodgers	147	574	108	170	38	8	12	1	85	57	2	37	8	-	7	7	.296	.367	.453
1949	30	Dodgers	156	593	122	203	38	12	16	0	124	86	12	27	17	-	8	22	.342	.432	.528
1950	31	Dodgers	144	518	99	170	39	4	14	1	81	80	14	24	10	-	5	11	.328	.423	.500
1951	32	Dodgers	153	548	106	185	33	7	19	0	88	79	6	27	6	-	9	10	.338	.429	.527
1952	33	Dodgers	149	510	104	157	17	3	19	0	75	106	8	40	6	-	14	16	.308	.440	.465
1953	34	Dodgers	136	484	109	159	34	7	12	0	95	74	5	30	9	-	7	12	.329	.425	.502
1954	35	Dodgers	124	386	62	120	22	4	15	0	59	63	4	20	5	4	7	13	.311	.413	.505
1955	36	Dodgers	105	317	51	81	6	2	8	0	36	61	5	18	6	3	3	8	.256	.378	.363
1956	37	Dodgers	117	357	61	98	15	2	10	0	43	60	2	32	9	2	3	9	.275	.382	.412
Career			G	AB	R	H	2B	3B	HR	GRSL	RBI	BB	IBB	SO	SH	SF	HBP	GIDP	AVG	OBP	SLG
10 Years			1,382	4,877	947	1,518	273	54	137	2	734	740	58	291	104	9	72	113	.311	.409	.474



**IN GOD
WE TRUST**

ALL OTHERS

**BRING
DATA**

Perchè Data Science?

Data Scientist: The Sexiest Job of the 21st Century

Meet the people who
can coax treasure out of
messy, unstructured data.

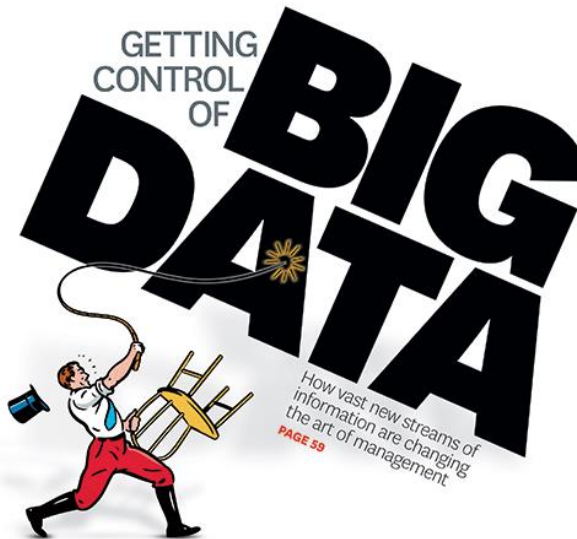
by Thomas H. Davenport
and D.J. Patil

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't

The shortage of data scientists is becoming a serious constraint in some sectors.

Harvard Business Review

OCTOBER 2012
46 The Big Idea
The True Measures
Of Success
Michael J. Mauboussin
84 International Business
10 Rules for Managing
Global Innovation
Kenley Wilson and Yves L. Doz
93 Leadership
What Ever Happened
To Accountability?
Thomas E. Ricks







your organization stores much of the information most critical to it in forms other than rows and columns. If answering your biggest questions is a “mashup” of several analytical data opportunities.

Much of the current enthusiasm focuses on technologies that include Hadoop (the most popular for distributed file system), open-source tools, cloud collaboration. While those are important at least as important are the people (and the mind-set) to put them to the front, demand has raced ahead.

Salaries by Company

Salaries in \$ (USD)

	Data Scientist Facebook (38 Facebook Data Scientists) \$ 135,359
	Data Scientist Twitter (15 Twitter Data Scientists) \$ 134,861
	Data Scientist Airbnb (14 Airbnb Data Scientists) \$ 117,229
	Data Scientist Microsoft (12 Microsoft Data Scientists) \$ 119,692

**Data scientist:
the next hot job**

Data Scientist

#1 JOB IN AMERICA

Cerca**181 risultati per data scientist**Ordina per **Pertinenza** ▾

Avanzata >

Tutto

Offerte di lavoro

Altro...

Parole chiave**Azienda****Titolo****Località****Paese****CAP****CRM Data Scientist**

lastminute.com group

Chiasso • 6 mag 2016

Simile

Visualizza ▾**Senior Data Scientist.**

NAGRA

Cheseaux, Switzerland • 5 mag 2016

Simile

Visualizza ▾**Senior Data Scientist**

Centralway Numbrs AG

Zürich, Switzerland • 25 apr 2016

Simile

Visualizza ▾**Senior Data Scientist - Think Big**

Think Big, A Teradata Company

Switzerland • 12 mag 2016

Simile

Visualizza ▾**Senior Data Scientist**

Credit Suisse

Zürich • 11 mag 2016

Simile

Visualizza ▾**Data Scientist**

Credit Suisse

Visualizza ▾

5 Aprile 2016 – GPU technology conference
Jensen H Huang CEO NVIDIA cita SUPSI-IDSIA tra i 10 pionieri nell'AI
www.gputechconf.com



Logos of the 10 pioneers in AI research:

- Berkeley
- Carnegie Mellon University
- Tsinghua University
- MIT
- Massachusetts Institute of Technology
- NYU
- Stanford University
- Université de Montréal
- University of Oxford
- IDSIA
- University of Toronto

Frameworks for
Multi-GPU Pascal

Large-scale Deep Learning

Reinforcement Learning

Unsupervised and Transfer
Learning

Natural Language
Understanding

Autonomous Driving

Medical Applications

**PIONEERS IN
AI RESEARCH**



Perche' studiamo data science

Negli ultimi anni abbiamo assistito a:

- *crescita esponenziale dei dati generati*
- *incremento considerevole dei dati registrati*

dovuta a:

- *riduzione drastica del costo di memorizzazione*
- *aumento significativo connettività*

In che ambiti?

industria

finanza

medicina

servizi

pubblica
amministrazione

vita quotidiana



I dati disponibili sono molto eterogenei per

- *origine*
- *contenuto*
- *rappresentazione*

Transazioni commerciali,
finanziarie, amministrative

percorsi di navigazione WEB

e-mail

*Testi e
ipertesti*

Esami medici

*Output di
sensori*

*risultati di test
clinici*



Un esempio pratico: retention nella telefonia mobile

Il responsabile marketing di un operatore di telefonia mobile si accorge che un numero crescente di clienti richiede la disattivazione del contratto per stipularne uno con un concorrente. In questi casi si parla di scarsa lealtà del cliente o anche di *customer attrition* o ancora di *churn*, fenomeno molto diffuso nel settore dei servizi.

Il responsabile marketing dispone di un budget che gli consente di condurre un'azione di retention su 200k clienti dei 4000k che ha attualmente.

Come sceglie quali dei 4000k clienti devono far parte dei 200k verso i quali attiverà un' *azione di retention* al fine di massimizzare l'efficacia della campagna di marketing?

Per individuare i 200k clienti si potrebbe procedere alla stima della probabilità di abbandono da parte di ogni cliente e rivolgere la campagna verso quei 200k clienti cui competono maggiori valori di tale probabilità.

Come si assegna il valore della probabilità di abbandono?

Altri esempi?

Previsione consumi

Segmentazione portfolio
clienti per azioni mirate

market basket analysis

Pianificazione delle
risorse aziendali

*Soluzioni Data-Driven a problemi di
riconoscimento di pattern*

Quali sono gli ingredienti?

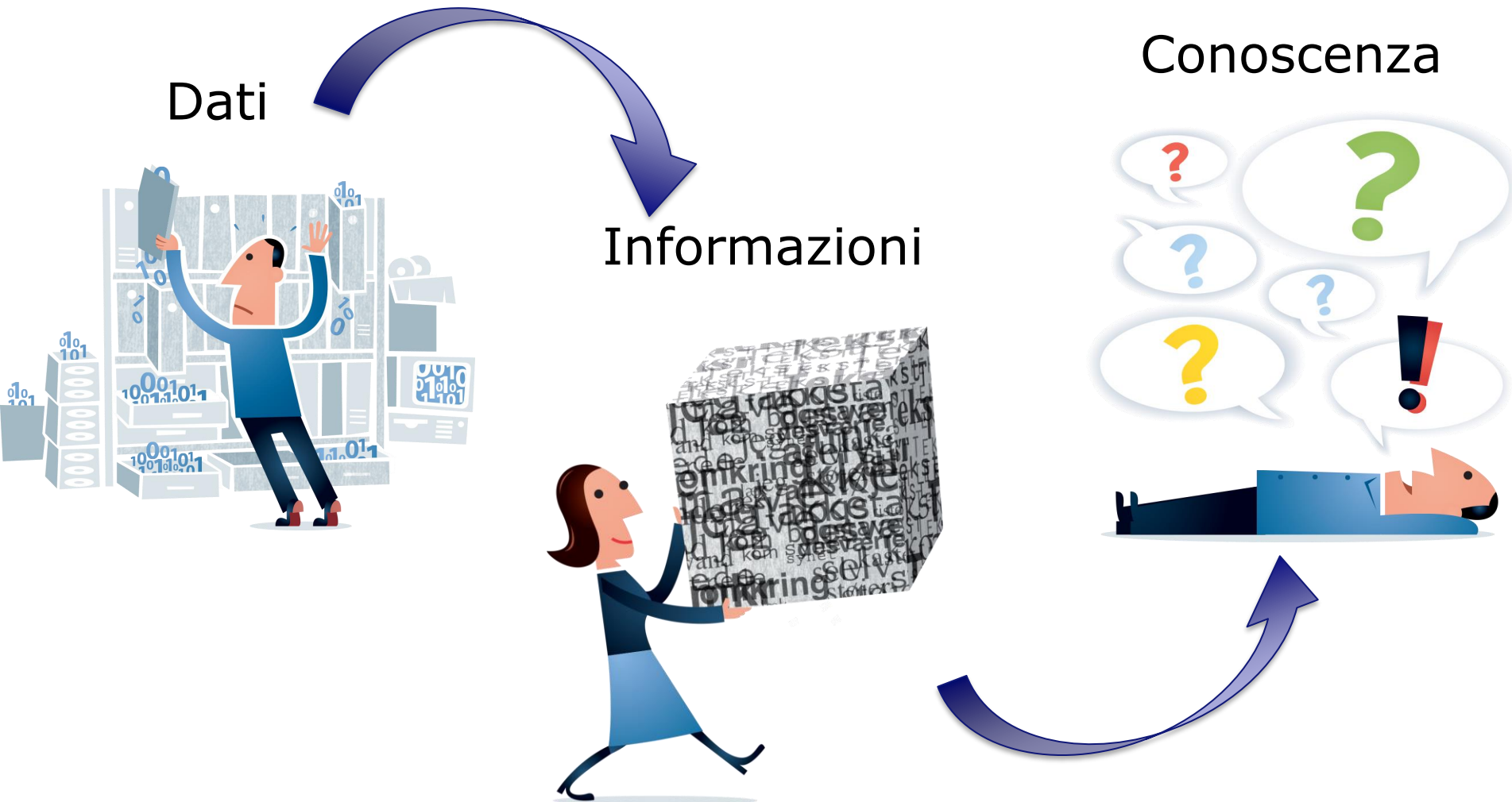
Domande giuste



Quali sono gli ingredienti?

Dati





Dati:

codifica strutturata di singole entità primarie e delle transazioni che coinvolgono due o più entità primarie.

Esempio: In un'azienda di grande distribuzione i dati fanno riferimento ai **clienti**, ai **punti vendita**, agli **articoli** mentre le transazioni commerciali sono descritte tramite gli **scontrini d'acquisto**.

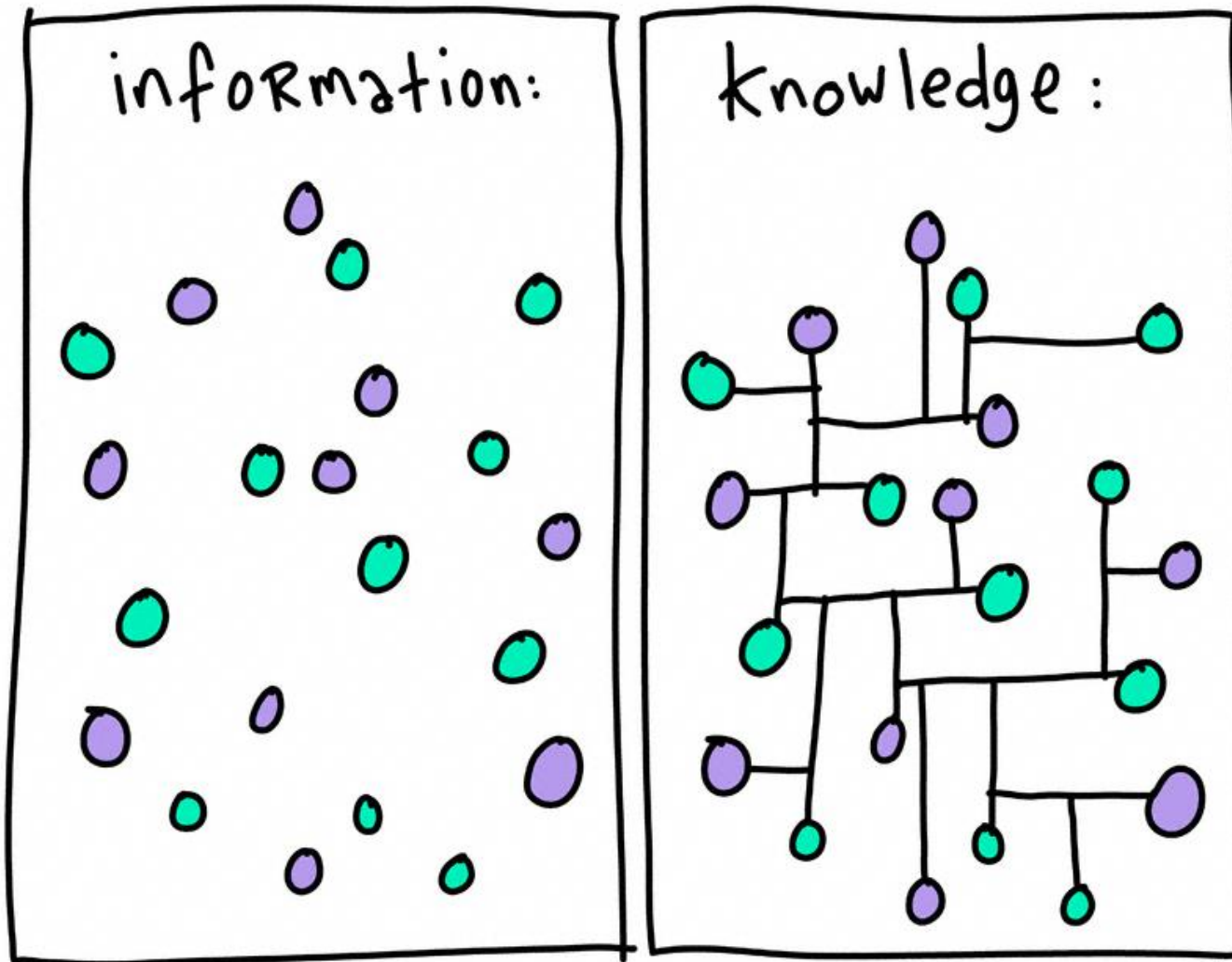
Informazioni:

risultato di operazioni di estrazione ed elaborazione compiute sui dati, hanno significato per chi le riceve in uno specifico contesto.

Esempio: Per un responsabile commerciale di un'azienda della grande distribuzione le informazioni sono, p.e., **la percentuale scontrini con importo superiore a 100 Euro sull'orizzonte temporale di una settimana, numero di possessori di carta-fedeltà che hanno ridotto più del 50% la spesa nell'arco di un mese, ...**

Conoscenza:

informazioni inserite in un contesto, arricchite da esperienza e competenze del *decision maker* per affrontare e risolvere problemi complessi.



@gapingvoid

Dati



Gestione dei dati

- Acquisizione: mettiamo le mani sui dati
 - Da dove vengono i dati?
 - Che aspetto hanno?
 - Come facciamo ad accedervi?
- Ingestione: facciamo in modo che i dati arrivino nel nostro sistema
 - Quanti dati arrivano?
 - Quanto in fretta?
 - Dove li mettiamo? Abbiamo abbastanza spazio su disco? Dobbiamo filtrarli?
- Trasformazione
 - In che formato sono i dati?
 - Come facciamo a trasformarli in un formato adatto all'analisi?

Esempio (discusso in aula)

- Un sistema di raccolta dati installato su un'autostrada e' dotato di un sensore per ogni corsia. Ogni sensore registra ogni passaggio di auto per ciascuna corsia, e salva le seguenti informazioni:
 - Timestamp (data e ora)
 - Auto, moto oppure camion
 - Velocita' del veicolo
- Che aspetto avranno i dati che arrivano dai sensori?
 - Cos'e' un' osservazione?
 - Quali sono gli attributi?
 - Quali sono i tipi di ciascun attributo?
- Possiamo trasmettere i dati dai sensori al nostro server tramite una connessione 3G? Abbiamo bisogno di una fibra ottica?
- Vorremo memorizzare un anno di dati, e poi elaborarli:
 - Ci bastera' una chiavetta USB?
 - I dati ci staranno nella RAM del mio laptop?
- A quali di queste domande potro' rispondere?
 - Quali sono gli orari di punta?
 - In quali orari e' piu' frequente che si formino delle code?
 - Quale percentuale di automobilisti usa il cellulare alla guida?
 - Quanti camion circolano la domenica?

Esercizio

- Una catena di supermercati ci fornisce giornalmente i dati provenienti dalle casse di tutti i punti vendita. Per ogni scontrino, ci viene fornito:
 - Timestamp
 - L'identificativo del cliente (se dotato di carta fedeltà)
 - Il totale dello scontrino
- Abbiamo inoltre il database dei clienti con le loro informazioni anagrafiche (sesso, età)
- Che aspetto avranno i dati?
 - Cos'è un'osservazione?
 - Quali sono gli attributi?
 - Quali sono i tipi di ciascun attributo?
- Possiamo trasmettere i dati da ogni supermercato al nostro server tramite una connessione 3G? Ci basta un piano economico da 250MB al mese?
- Vorremo memorizzare un anno di dati, e poi elaborarli:
 - Ci basterà una chiavetta USB?
 - I dati ci staranno nella RAM del mio laptop?
- A quali di queste domande potrò provare a rispondere, e a quali no?
 - In quali orari del giorno ci sono più clienti?
 - Quanti clienti perderemmo se decidessimo di chiudere un'ora prima? Quanti clienti guadagneremmo chiudendo un'ora dopo?
 - Possiamo trasmettere all'interno dei supermercati degli annunci promozionali che sappiamo essere particolarmente efficaci sui maschi nella fascia d'età 20-30 anni. In che orario della giornata è meglio trasmetterli?
 - Dobbiamo assumere più cassiere?
 - Possiamo pubblicizzare un'auto di lusso in uno dei nostri punti vendita: quale punto vendita è frequentato da persone con alto reddito?

Esercizio

- Una squadra professionistica di pallavolo ha implementato un sistema di raccolta dati in uso durante le partite della stagione. Per ogni tocco di palla di uno dei propri giocatori, una persona dotata di un apposito software registra:
 - L'istante esatto del tocco (data e ora, con precisione al millisecondo)
 - Il numero del giocatore
 - Il tipo di tocco (battuta, ricezione, alzata, schiacciata, muro)
 - Una valutazione da 1 (pessimo) a 5 (ottimo)
- Che aspetto avranno i dati?
- Quanto spazio occuperanno i dati di una partita, approssimativamente?
- L'allenatore vuole identificare i punti di forza e i punti deboli di ciascuno dei suoi giocatori. Può farlo? Come?
- L'allenatore vuole verificare se un innovativo regime di allenamento mirato a migliorare la battuta è efficace. Come può farlo?

Esempio: <http://xkeys.com/utilization/sportsdatarecording.php>

Esercizio (speculativo)

Vuoi sviluppare un sistema che possa essere utile per un problema che conosci bene (ad esempio in ambito lavorativo). Descrivi:

- Ci sono delle domande a cui avrebbe senso rispondere analizzando dei dati?
- Quali dati sarebbe necessario acquisire? Come pensi sia possibile acquisire questi dati (ad esempio: dei sensori; degli osservatori; dei software di analisi video)
- Discuti brevemente come elaboreresti i dati