SI PUÒ VIVERE SENZA SCIENZA?

BIGDATA e metodo scientifico

Giovanni Pistone
www.giannidiorestino.it
in collaborazione con
Francesca Dell'Orto

DE CASTRO STATISTICS

Collegio Carlo Alberto

Roma, 3 marzo 2017

Sunto

Secondo alcuni entusiasti del BigData e dei relativi algoritmi di apprendimento statistico, la scienza moderna non ha bisogno di appoggiarsi su modelli teorici, ma anzi è in grado di ricavare a posteriori la teoria che è necessaria sulla base dello studio delle correlazioni tra le variabili di interesse. Sia la pratica della ricerca scientifica che l'analisi di cosa effettivamente fanno gli algoritmi di apprendimento statistico mostrano quanto sia grossolanamente esagerata questa affermazione. Questi analisti puntano però il dito in una direzione che non può essere ignorata concentrando la critica sul dito che indica. I sistemi di acquisizione automatica di misure, la rete, e le grandi industrie come Google forniscono effettivamente un servizio di straordinario impatto sulla società e anche sulla ricerca scientifica, servizio che è certamente indispensabile comprendere. A partire dalla nozione di modello statistico e di modello teorico, analizziamo il contenuto concettuale di alcuni indiscutibili successi dell'algoritmica moderna, tra cui l'ottimizzazione con algoritmi genetici e le reti neurali per il riconoscimento di immagini. In conclusione, consideriamo due ipotesi di interpretazione dei limiti degli algoritmi di apprendimento, cioè il carattere di base delle credenze ottenute e la loro conseguente incapacità di produrre autentici risultati nella direzione del supporto all'azione e allo sviluppo della conoscenza.

Piano di lavoro

1. Assumiamo lo stesso punto di partenza di molti, cioè un articolo pubblicato da Chris Anderson nel 2008 sulla rivista *Wired*. Osserviamo che Anderson pone al centro, giustamente, la nozione di modello. In presenza di BigData i modelli si studiano attraverso tecniche di riduzione delle dimensioni.
2. L'esplicito riferimento di Anderson è una grande impresa industriale contemporanea, cioè Google. Questa industria produce con tecniche di intelligenza artificiale e poi distribuisce strumenti di conoscenza di un tipo particolare, molto interessanti, ma che non sono orientati alla scienza. Esaminiamo in particolare gli algoritmi genetici e le reti neurali.
3. Individuata l'effettiva direzione di marcia della tecnologia in questione, passiamo a considerare alcune ipotesi sulle implicazioni sul piano epistemico e filosofico. Consideriamo due tesi riguardo al carattere delle conoscenze acquisite.

CHRIS ANDERSON

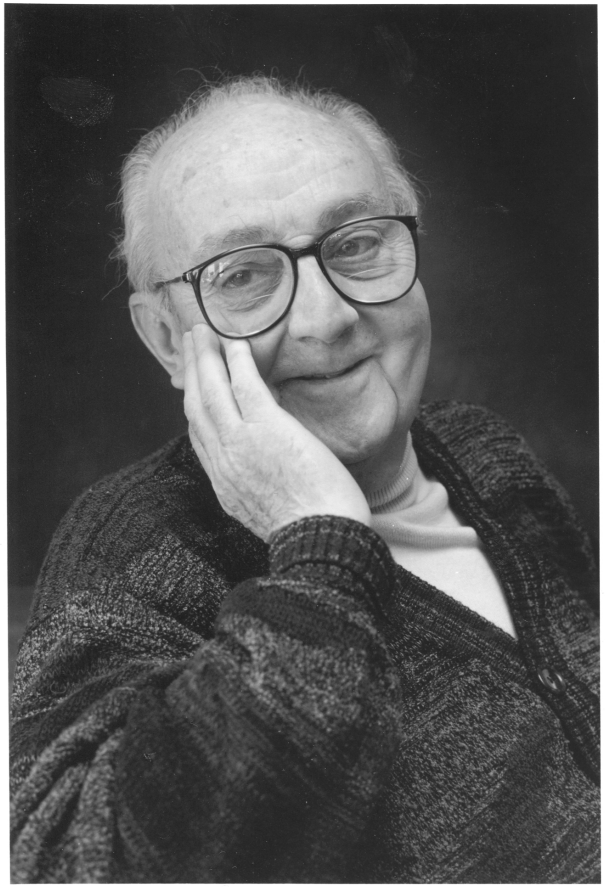
Wired Jun 23, 2008. *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*

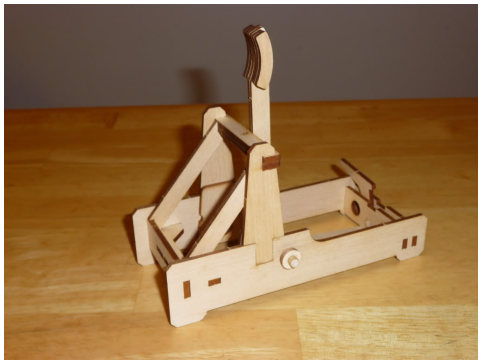
"**All models are wrong**, but some are useful." So proclaimed statistician George Box 30 years ago, and he was right. But what choice did we have? Only models, from cosmological equations to theories of human behavior, seemed to be able to consistently, if imperfectly, explain the world around us. Until now. Today companies like Google, which have grown up in an era of massively abundant data, don't have to settle for wrong models. Indeed, they don't have to settle for models at all.

In evidenza

- Tutti i modelli sono sbagliati, ma alcuni sono utili.
- Così proclamava lo statistico George Box 30 anni fa, e aveva ragione.
- Ma quale altra scelta avremmo avuto? Solo i modelli, da quelli cosmologici alle teorie del comportamento umano, sembravano essere in grado di spiegare in modo coerente, anche se imperfetto, il mondo intorno a noi.
- Fino ad ora. Oggi industrie come Google, che sono cresciute in un'epoca di massiccia abbondanza di dati, non devono accontantatsi di modelli sbagliati. In effetti, non devono per nulla eccontentarsi di modelli.

Ronald Fiscer e George Box

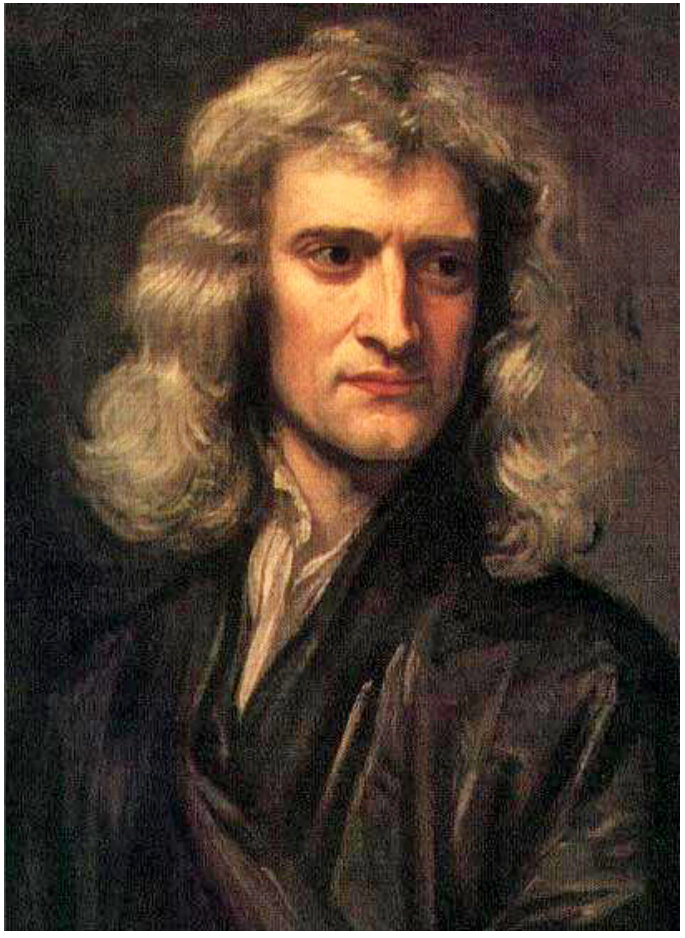


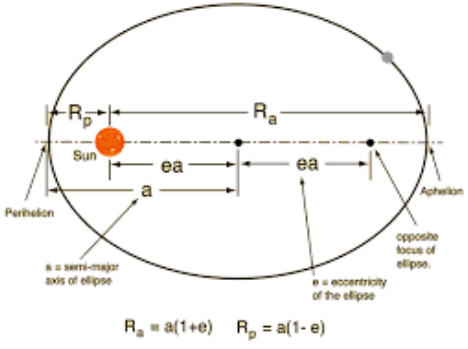


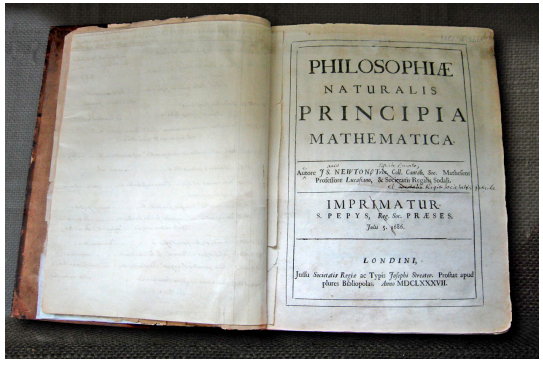
Protocollo statistico-sperimentale:

1. definire quali sono le variabili rilevanti del problema sulla base dell'opinione degli esperti;
2. proporre un modello matematico parsimonioso, cioè semplice, che lega in una espressione matematica le variabili in questione e un numero (limitato) di parametri incogniti;
3. pianificare un esperimento facendo variare in modo sistematico le variabili controllabili e misurare quelle di risposta in corrispondenza di ogni configurazione delle variabili controllabili;
4. stimare il valore di quei parametri che adattano meglio il modello alle risposte osservate;
5. leggere sul modello identificato la risposta di interesse;
6. fornire una valutazione dell'errore dedotta dal protocollo sperimentale utilizzato.

Keplero o Newton?



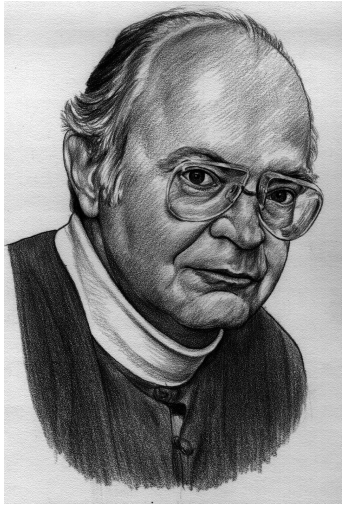


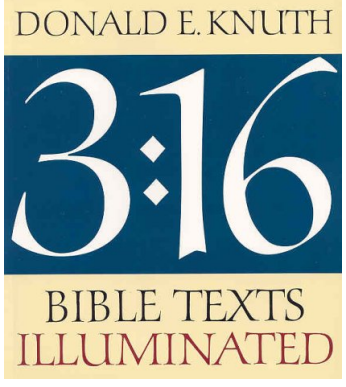


Riduzione della dimensione:

- indici (medie . . .);
- campionamento;
- proiezione;
- proiezione casuale;
- calcolo delle distribuzioni;
- convoluzione
- . . .

Un campionamento di Donald Knuth





Genesi 3,16 Alla donna disse: Moltiplicherò i tuoi dolori e le tue gravidanze, con dolore partorirai figli. Verso tuo marito sarà il tuo istinto, ma egli ti dominerà.

⋮

Isaia 3,16 Dice il Signore: Poiché si sono insuperbite le figlie di Sion e procedono a collo teso, ammiccando con gli occhi, e camminano a piccoli passi facendo tintinnare gli anelli ai piedi,

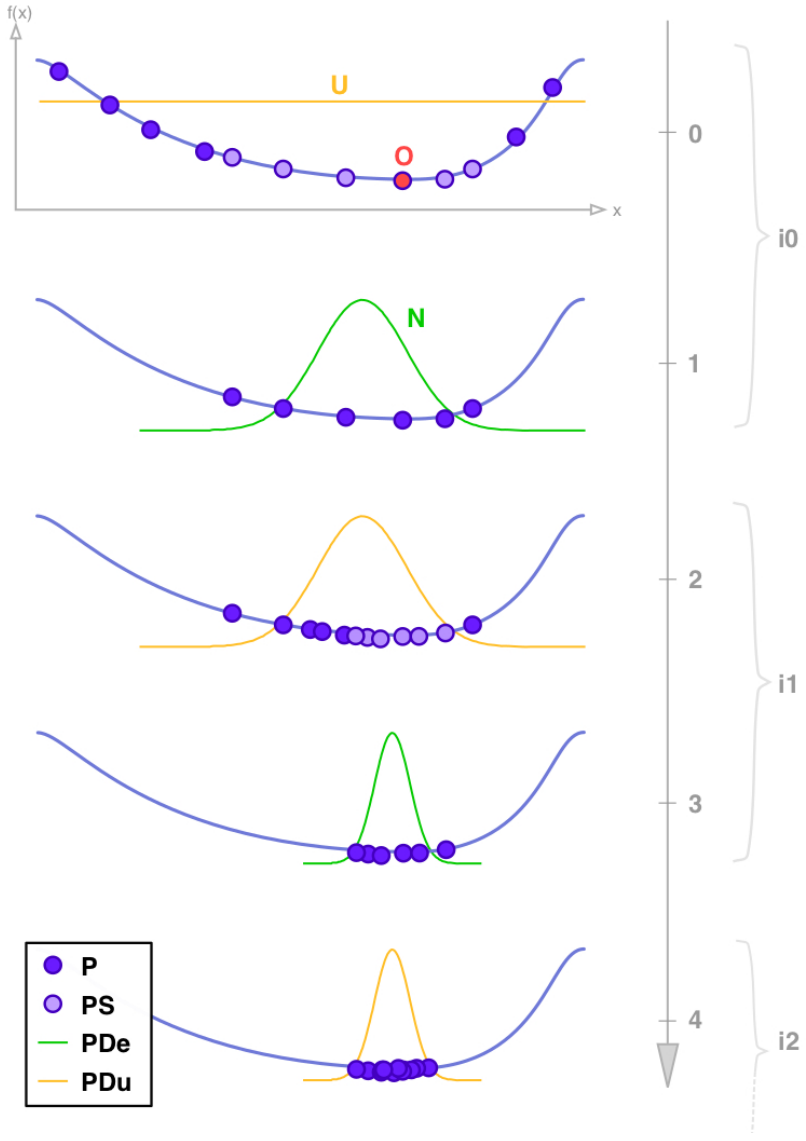
⋮

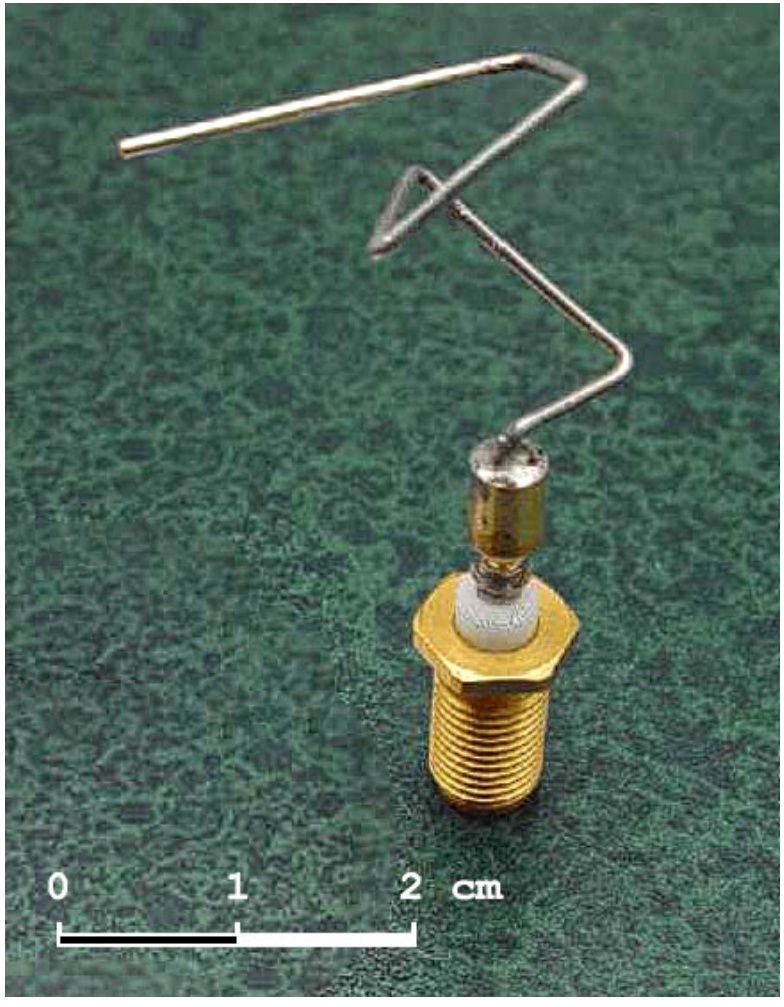
Giovanni 3,16 Dio infatti ha tanto amato il mondo da dare il suo Figlio unigenito, perché chiunque crede in lui non muoia, ma abbia la vita eterna.

⋮

Apocalisse 3,16 Ma poiché sei tiepido, non sei cioè né freddo né caldo, sto per vomitarti dalla mia bocca.

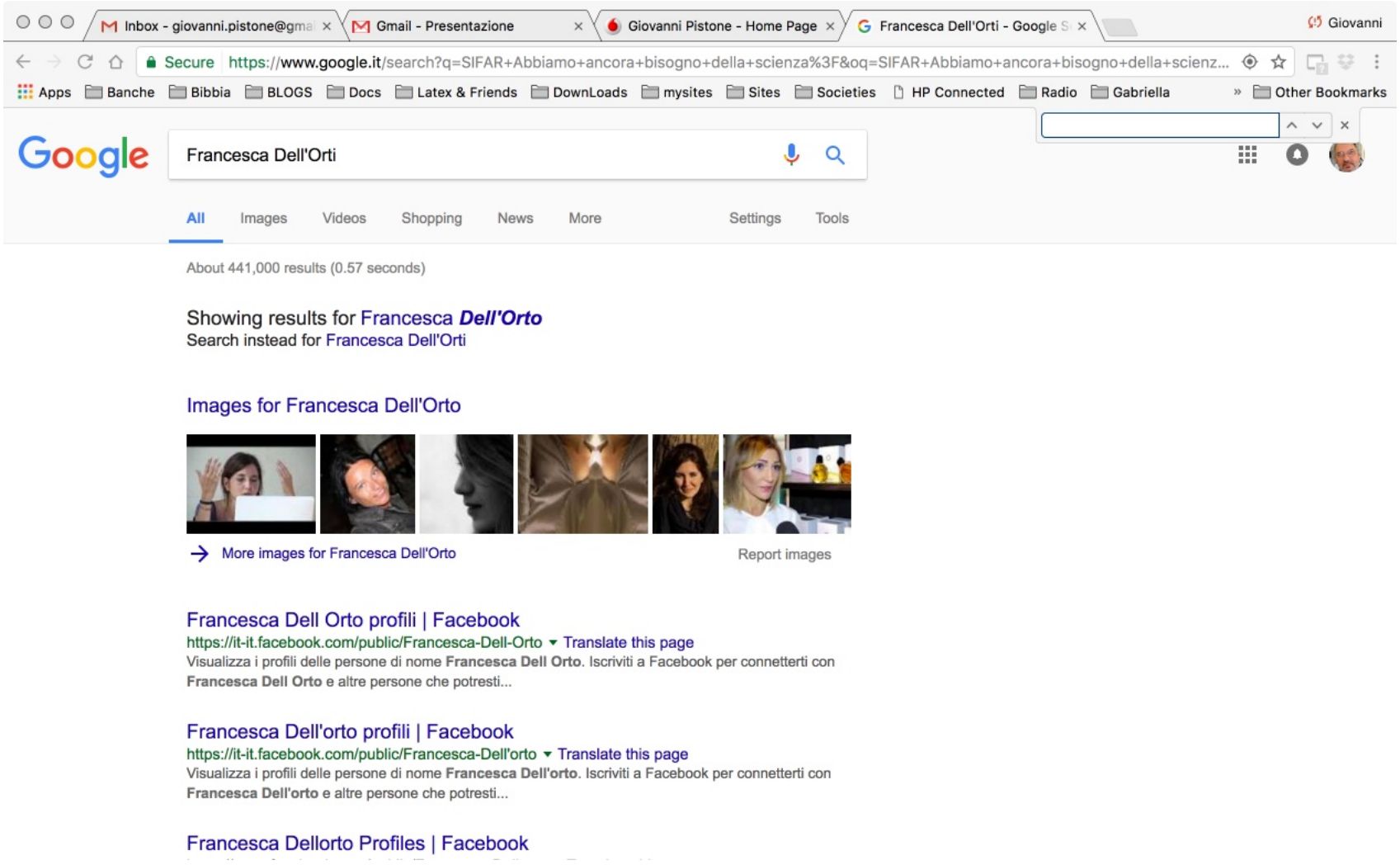
Algoritmi evolutivi: EDA



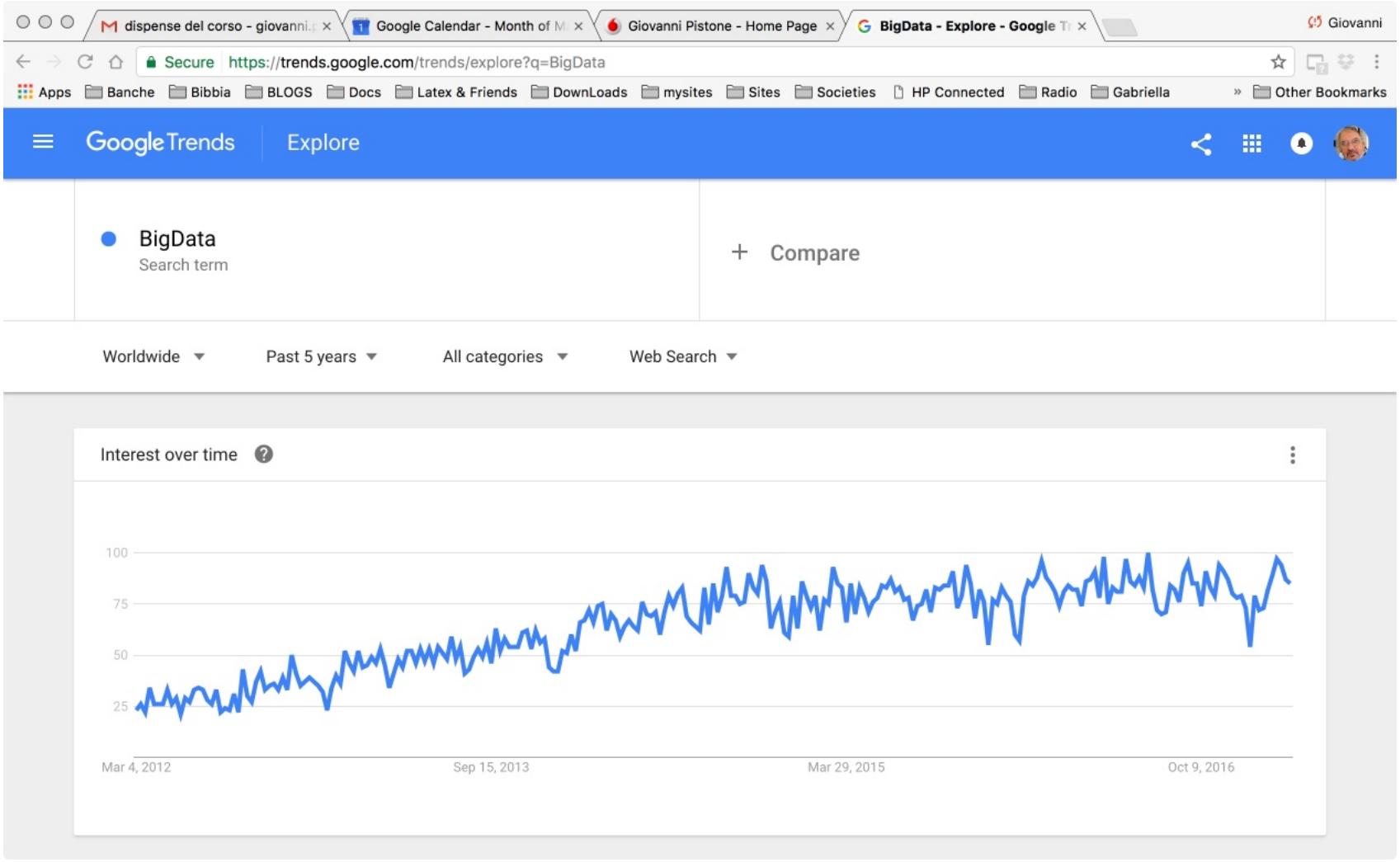


Da Wikipedia

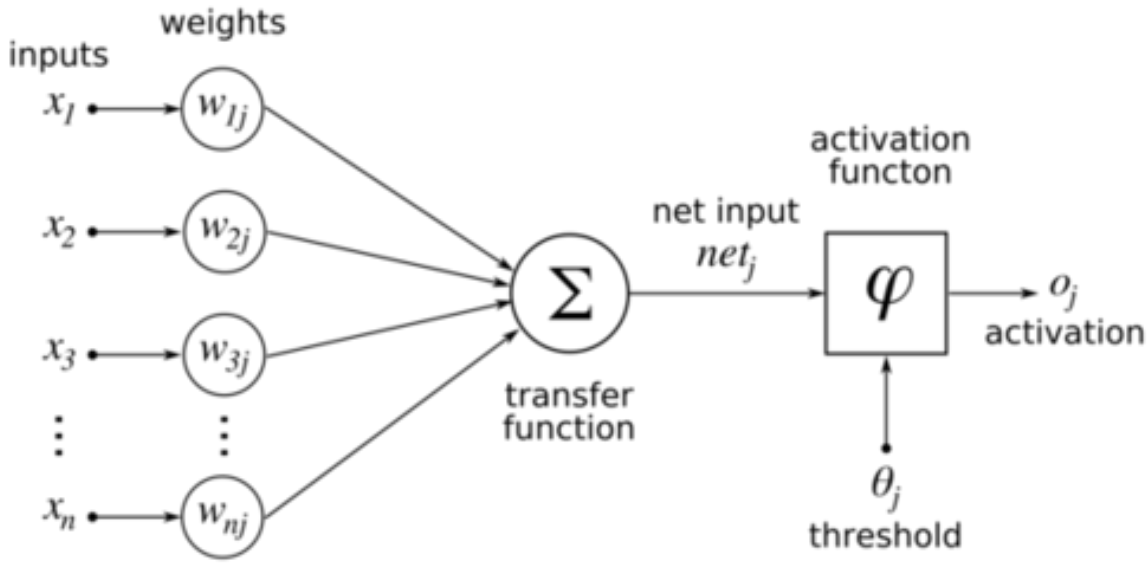
Controllo ortografico



Statistica delle ricerche

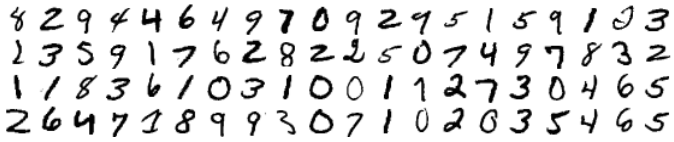


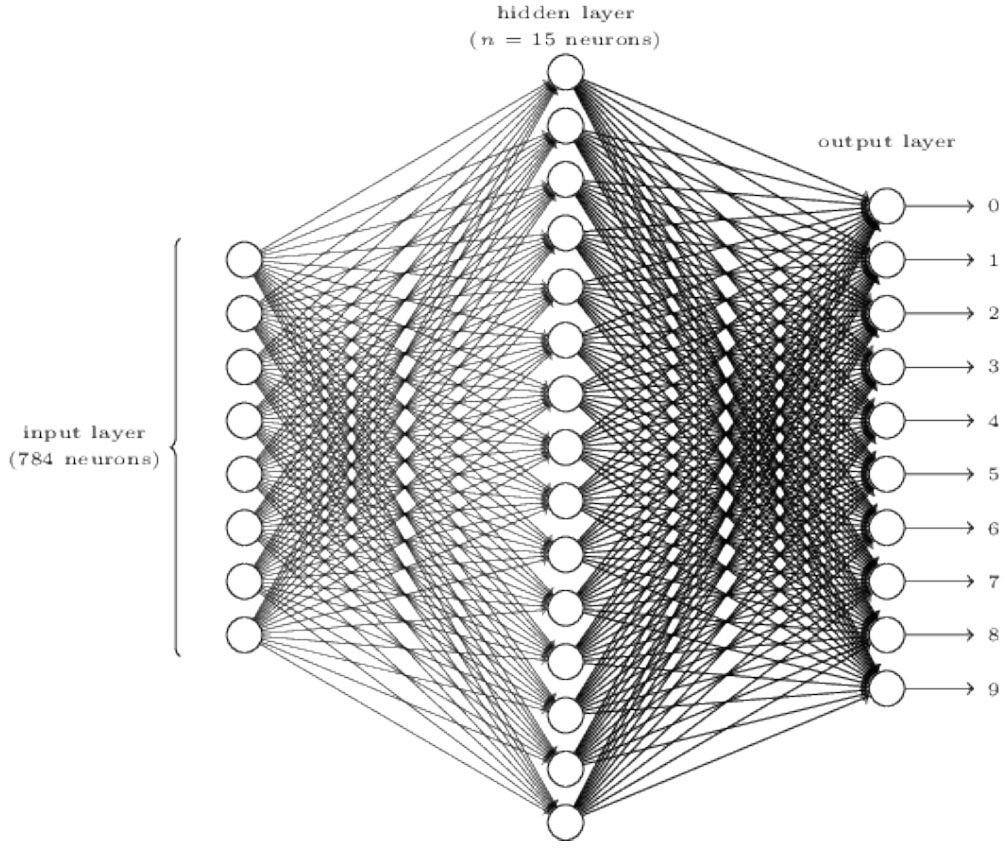
Percettrone di Frank Rosenblatt (1957)



$$o_j = \phi\left(\sum_{k=1}^{n} w_{k,j} x_k ; \theta\right)$$

Rete neurale

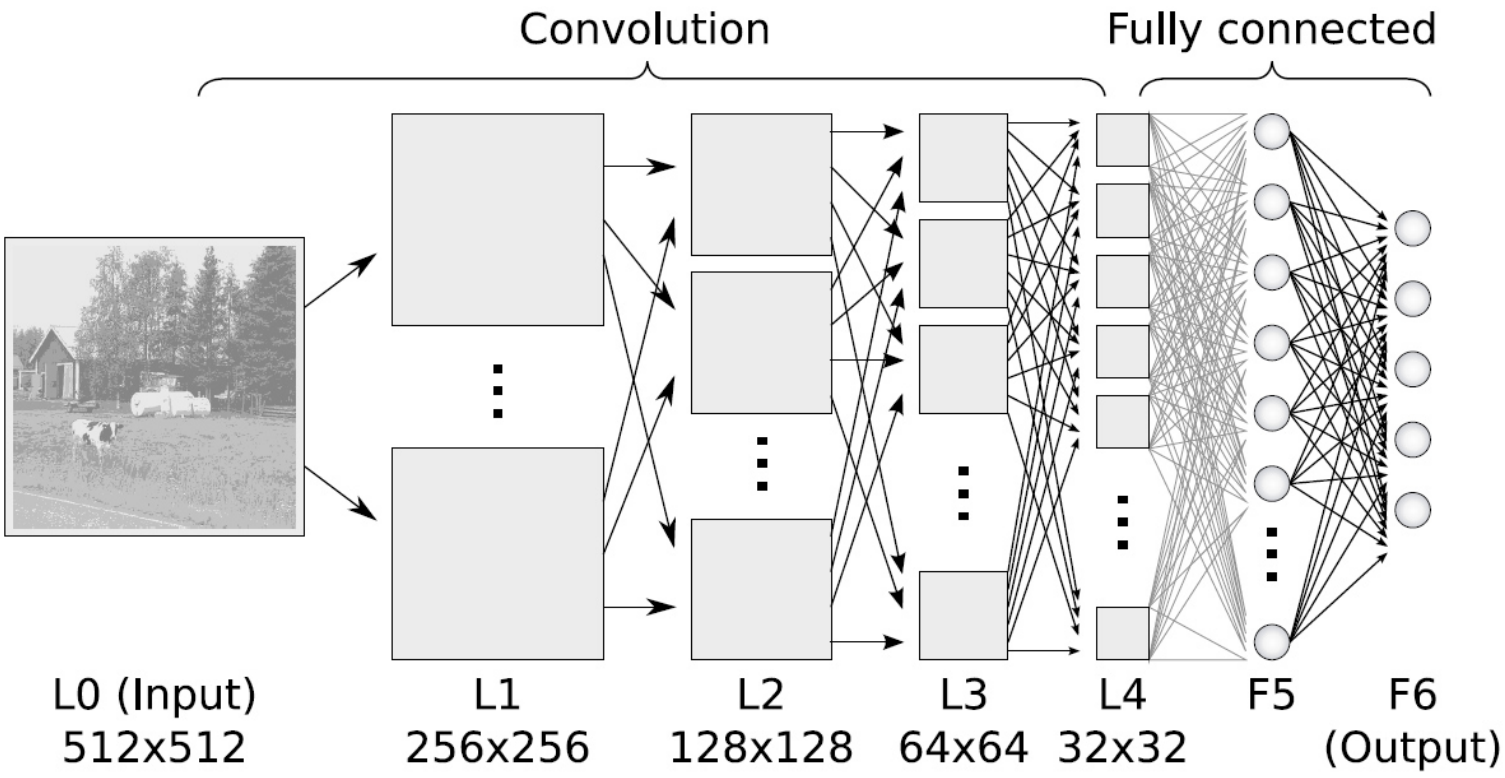




$$28 \times 28 = 784 \rightarrow$$

Cfr: `http://yann.lecun.com/exdb/mnist/`

Deep learning



Deep learning = rete neurale con molti strati

Image: `https://www.ais.uni-bonn.de/deep_learning/images/Convolutional_NN.jpg`

Risposte a Anderson

- non è vero che Google non usa modelli a priori, perché in effetti li usa in molti degli algoritmi interni;
- non è vero che Google deduce modelli a posteriori dallo studio delle correlazioni statistiche su grandi masse di dati, perché non lo fa, essendo per altro questa impresa impossibile.

Dice il redattore CADE METZ su *WIRED: BUSINESS* 11.16.2015:

GOOGLE OPEN-SOURCING TENSORFLOW SHOWS AI'S FUTURE IS DATA

Ma bisogna considerare:

- la rilevanza per la ricerca della disponibilità di dati numerosi e di pubblico dominio e di algoritmi di analisi;
- la nascita di una ideologia associata al Big Data e le sue implicazioni su larga scala;
- gli effetti sul comportamento e sulla coscienza di chi vive a contatto con gli oggetti tecnici contemporanei, prima di tutto la Rete.

Prima tesi:

- campionamento, algoritmi genetici, reti neurali e altri algoritmi simili ispirati da modelli di cervello umano, sono concepiti e destinati a simulare alcune, ma non tutte, le facoltà epistemiche umane;
- questi algoritmi simulano la formazione di credenze di base ricavate da dati di osservazione elaborati e di un numero limitato di leggi logiche o matematiche;
- Le credenze di base, se certificate da un meccanismo di acquisizione funzionante e diretto verso la verità, possono essere strutturate in conoscenza, cioè in credenze non basiche. Una struttura particolarmente coerente e vera di conoscenze costituisce la scienza vera e proria.

Seconda tesi:

- le credenze non sono solo brandelli di conoscenza, ma sono vere e proprie propensioni all'agire;
- l'oggetto tecnico rete ha una specifica relazione sulla concezione corrente dei rapporti temporali passato-futuro ovvero memoria-protensione (anticipazione del futuro);
- una protensione puramente basata sui dati e sulla loro elaborazione in credenze non può rendere conto dell'emergenza del nuovo, non solo nel senso evolutivo, ma anche nel semplice senso di scoperta dell'esistente ignoto e dunque inatteso;
- questi schemi sono sostanzialmente simmetrici nella direzione del tempo, e non possono in nessun modo rendere conto di quella che i fisici chiamano la *freccia del tempo*.
- forecasting $\rightarrow$ nowcasting

Messaggio da asporto

Nulla è più pratico di una buona teoria