

Scienza dei Dati e Intelligenza Artificiale nella Scuola Secondaria

10 febbraio 2020

Carlo Batini
Università di Milano-Bicocca

Le Basi della Scienza dei dati: concetti e relazioni
con la Intelligenza Artificiale e con altre discipline

Da La Repubblica di lunedì 31 gennaio 2020

LA STRATEGIA DI BRUXELLES

Dati e intelligenza artificiale

La sfida europea a Usa e Cina

Dopo il Green Deal
un nuovo piano Ue
da 200 miliardi
per un hi-tech più etico

dal nostro corrispondente
Alberto D'Argenio

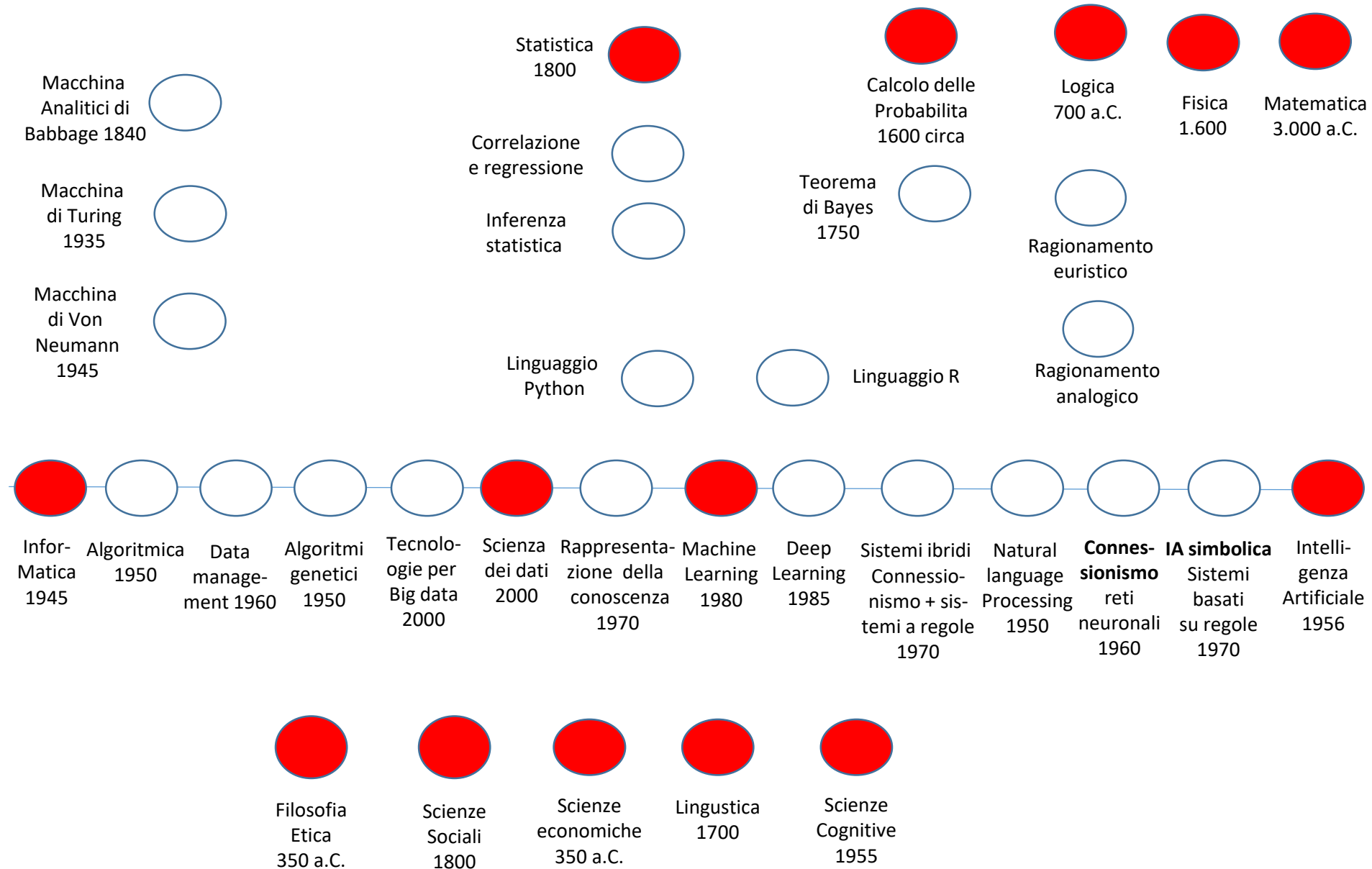
BRUXELLES – L'Europa sfida Stati Uniti e Cina su Intelligenza artificiale e Big data mettendo sul tavolo 200 miliardi di euro in dieci anni. Punta a essere il primo continente a

I numeri

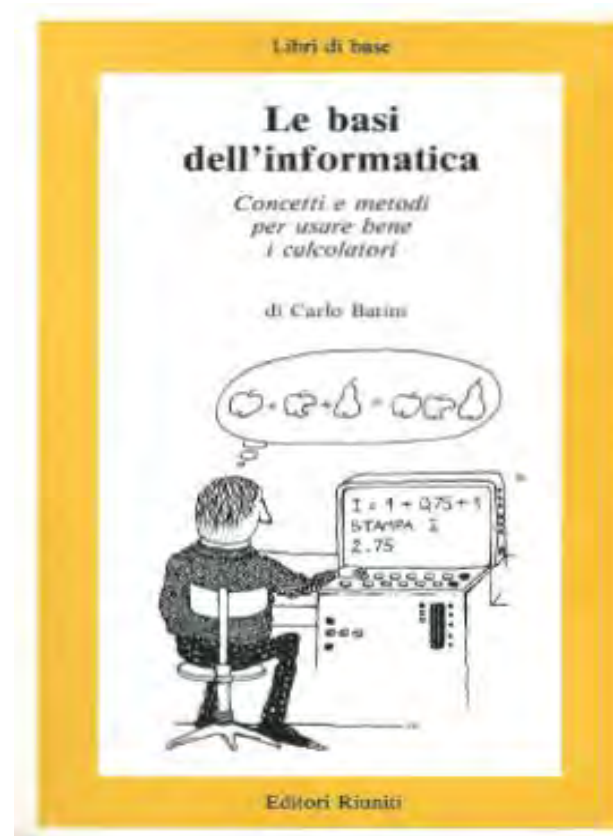
per l'Unione», scrive la Commissione. E allora l'Europa punta a trasformarsi in un grande hub globale fondato su regole etiche. Sul piatto 600 milioni per il cloud europeo e un miliardo per la costruzione di nuove app. Se oggi l'80% dei dati globali è detenuto dalle piattaforme americane e cinesi, grazie al diffondersi di "Internet delle cose" – elettrodomestici, auto e macchinari connessi – entro il 2025 sempre l'80% dei dati sarà contendibile. È proprio giocando sul passaggio da "Internet delle persone" a "Internet delle cose" che l'Europa cercherà di equilibrare la

gislazione europea. Una regolamentazione da applicare quanto meno ai settori ritenuti ad alto rischio, ovvero nei campi in cui la vita umana dipende da un algoritmo: salute, ordine e sicurezza pubblica, gestione del traffico e guida autonoma, aeroporti, ospedali, identificazione di massa. In questi settori l'intelligenza artificiale dovrà rispettare i diritti fondamentali riconosciuti dalla Ue, la privacy e la confidenzialità delle informazioni carpite all'umano (che dovrà sempre essere informato che sta interagendo con una macchina). Inoltre il funzionamen-

Mappa dei saperi nella Scienza dei dati e nella Intelligenza Artificiale



35 anni fa - Il libro “Le basi dell’informatica”



Scaricabile da <http://hdl.handle.net/10281/97703>

Corso e Libro “Le basi della Scienza dei Dati”



50

I LINCEI PER UNA NUOVA DIDATTICA NELLA SCUOLA: UNA RETE NAZIONALE
POLO DI MILANO - a.s. 2019/2020

Titolo corso: SCIENZA DEI DATI e INTELLIGENZA ARTIFICIALE NELLA SCUOLA SECONDARIA

Referente: Prof. Carlo Batini, Università di Milano Bicocca

Il documento presenta il Corso 2019/2020 per la nuova didattica nella Scuola organizzato dal Polo di Milano della Fondazione Lincei per la Scuola insieme all'Istituto Lombardo Accademia di Scienze e Lettere, la Fondazione Tronchetti Provera e la Università di Milano Bicocca.

Partendo dal tema della innovazione digitale, il corso si concentra in particolare sulle nascenti Scienza dei dati e Intelligenza Artificiale. La Scienza dei dati ha dato luogo negli ultimi anni a molte proposte di percorsi universitari in diversi paesi nel mondo, e in diverse Università italiane, l'Intelligenza Artificiale si sta ora affermando con analoghi percorsi formativi, che si stanno propagando verso la Scuola secondaria e in alcune esperienze verso la Scuola primaria.

Il corso intende sperimentare la erogazione di contenuti che potranno in futuro essere insegnati in Italia nella Scuola secondaria e negli Istituti tecnici superiori.

Come accade per molte scienze nella loro fase di formazione, non esiste ancora una visione condivisa su cosa si intenda per Scienza dei dati e su quali paradigmi e concetti essa si basi; siamo perciò ancora in fase nascente. Tuttavia, diversi fenomeni che tutti noi percepiamo nella nostra vita e nella evoluzione della ricerca, delle società e dei mercati, segnalano che i dati digitali stanno diventando un elemento di grande rilevanza e rapida trasformazione, e che attorno ai dati digitali si stanno formando contenuti, analisi, itinerari formativi che è urgente trasformare in un corpo di conoscenza da trasmettere e condividere.

Possiamo definire la Scienza dei dati come l'insieme dei modelli, metodologie, linguaggi, tecniche, e delle loro applicazioni, che permettono di elaborare, analizzare, ragionare su un vasto insieme di tipologie di dati digitali, come i dati tabellari, i testi, le immagini, essendo in grado di:

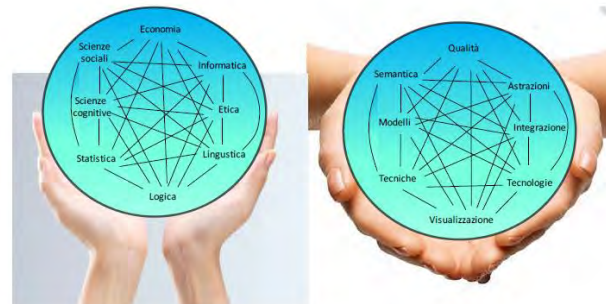
- ricostruire e modellarne il significato attraverso linguaggi concettuali;
- applicare tecniche per la valutazione del livello di veridicità e qualità,
- integrare dati eterogenei riconciliandone le differenze,
- interrogare e analizzare i dati mediante linguaggi e modelli per estrarne conoscenza
- applicare tecniche e modelli statistici e basati su apprendimento per costruire modelli decisionali, interpretativi, predittivi e prescrittivi
- visualizzare i dati digitali per comprendere meglio la natura dei dati e i risultati delle analisi
- risolvere problemi e assumere decisioni complesse utilizzando tecniche basate sulla Statistica e sul Machine Learning.
- comprendere l'impatto sulla economia e sulla società del fenomeno dei dati
- analizzare i corpi giuridici sviluppati dalle istituzioni pubbliche in tema di dati
- comprendere i principi delle scienze cognitive che presiedono all'uso consapevole dei dati nella nostra vita.
- affrontare i temi etici che nascono dall'uso dei dati.

La Scienza dei dati sta sviluppando metodi, tecniche, linguaggi, modelli che si ispirano a diverse altre Scienze. Anzitutto la Statistica, la Matematica e la Informatica, che nell'ambito della Statistica inferenziale e del Machine Learning forniscono potenti strumenti per la analisi dei dati. Accanto ad esse, la Economia è profondamente influenzata dai dati digitali, che

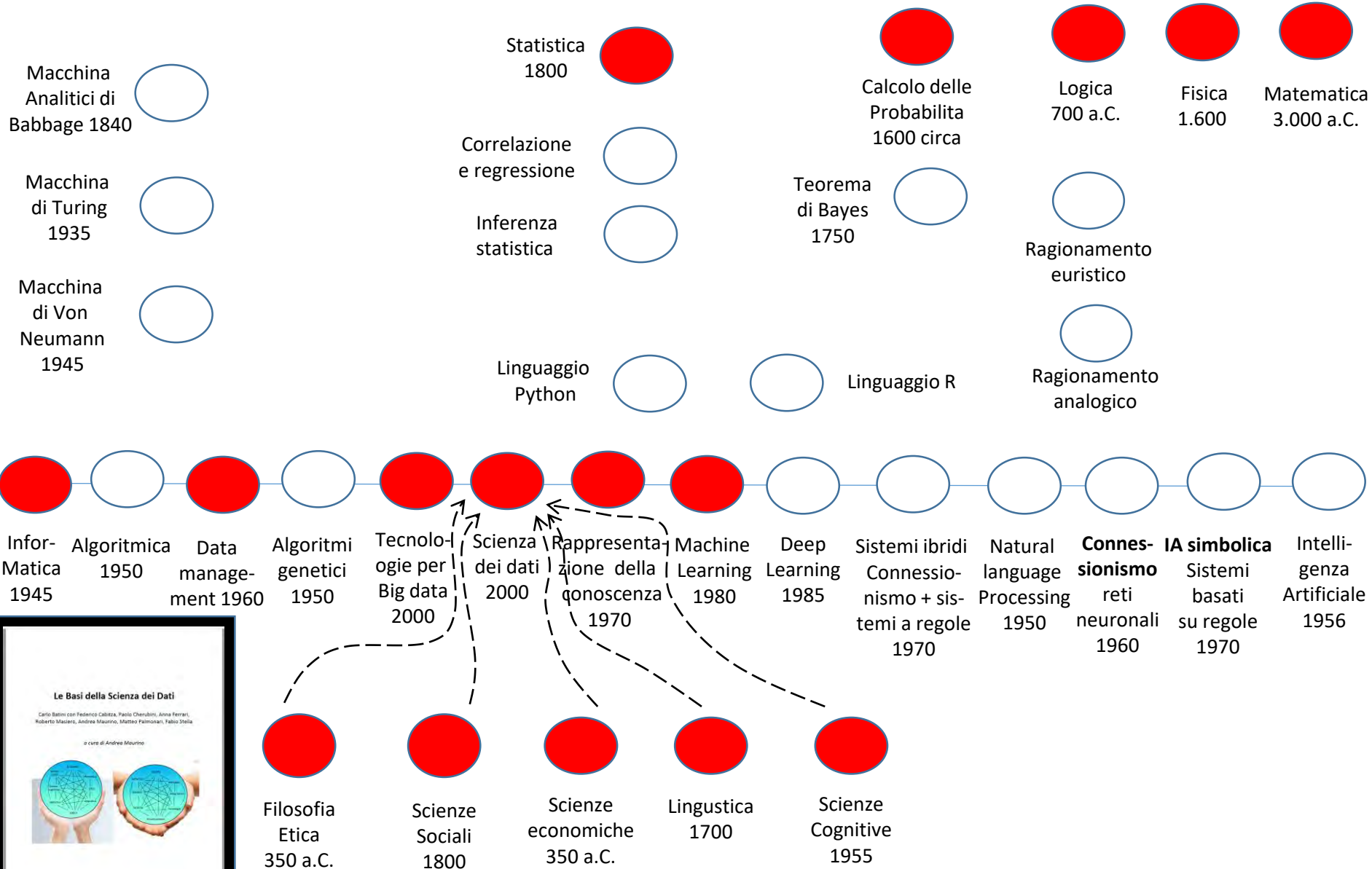
Le Basi della Scienza dei Dati

Carlo Batini con Federico Cabitza, Paolo Cherubini, Anna Ferrari, Roberto Masiero, Andrea Maurino, Matteo Palmonari, Fabio Stella

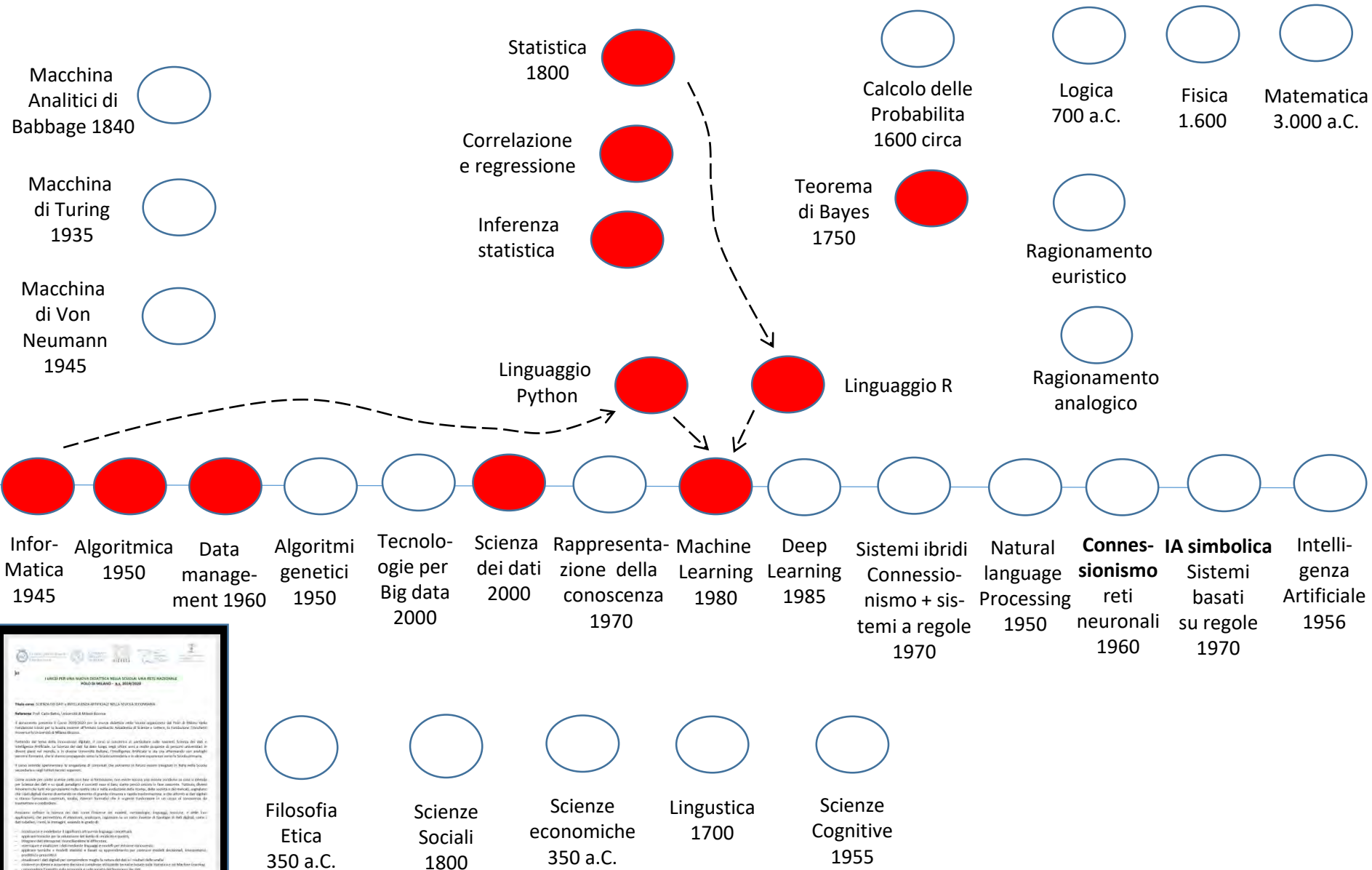
a cura di Andrea Maurino



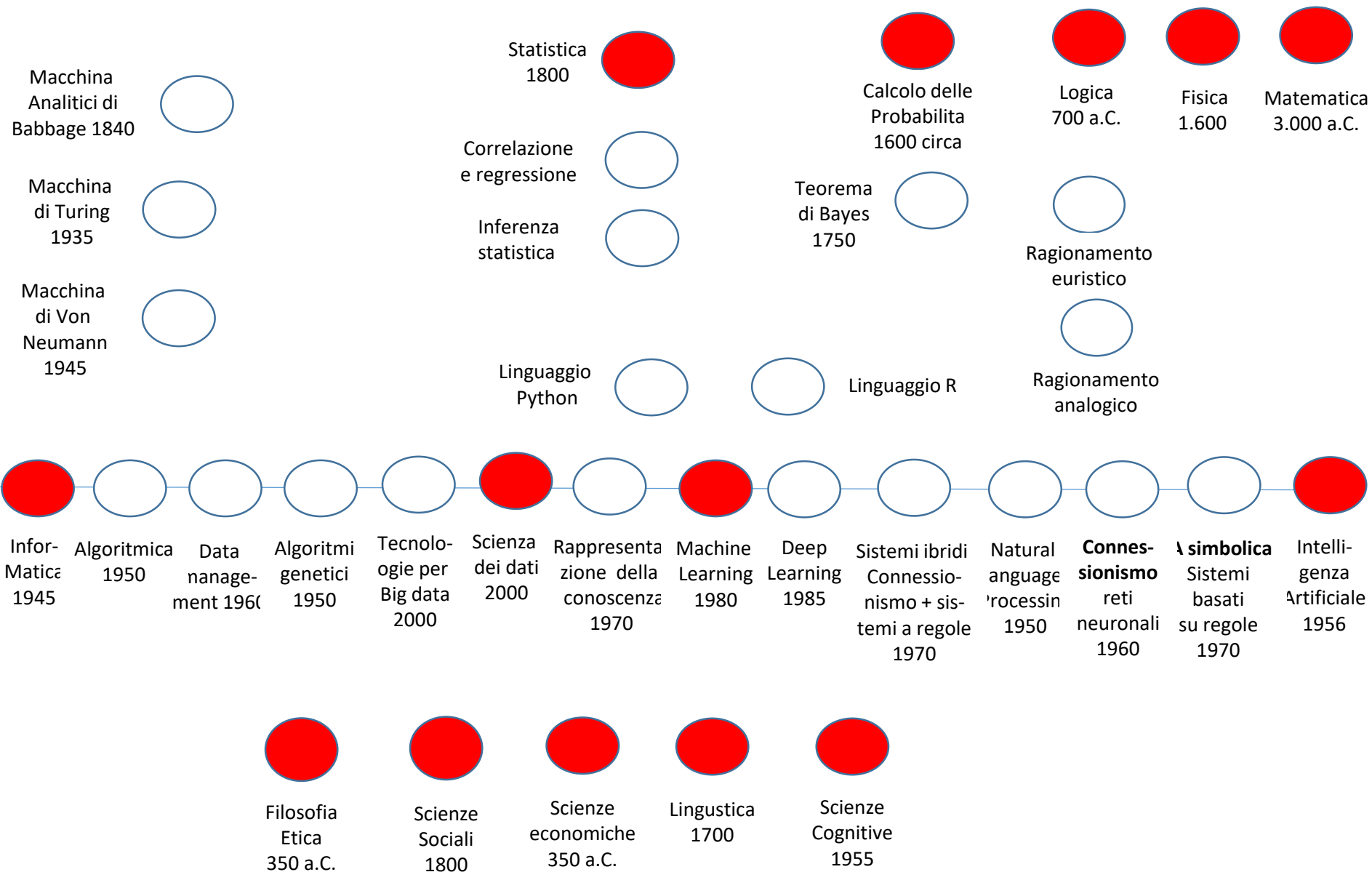
Mappa dei saperi nel Libro



Mappa dei saperi nel Corso



Da qui in poi



Questo Seminario introduttivo

Dati, informazioni, conoscenza
che differenza c'è?

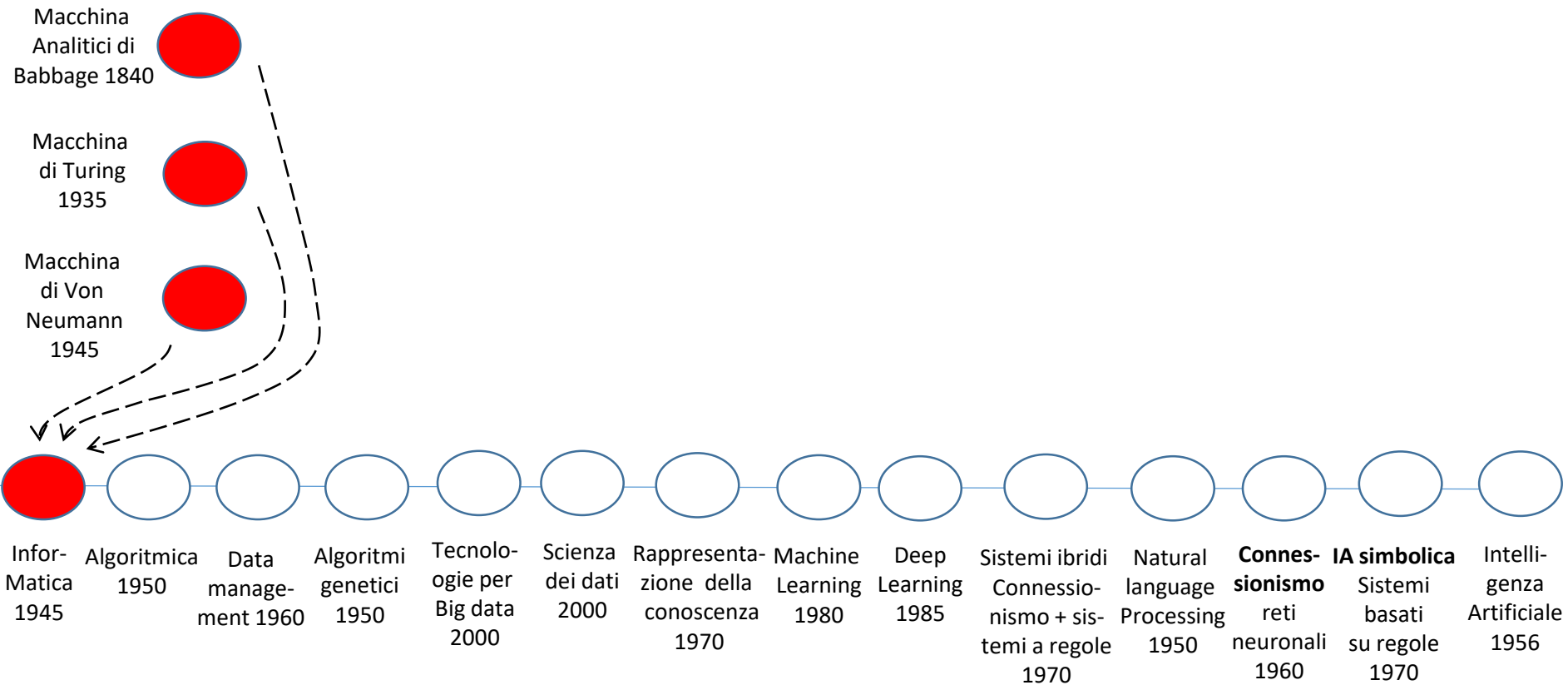
Dati, informazioni, conoscenza



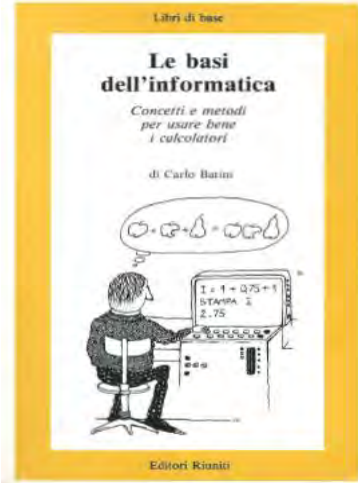
- 37.5 e' un **dato**
- Se io so (**conoscenza**) che
 1. Il termometro serve per misurare la temperature e
 2. Quel termometro misura la temperatura in gradi Celsius allora posso dire che 37.5 è una temperatura Celsius (**informazione**)
- Se poi so che
 3. la soglia per la febbre è circa 37 (**conoscenza**)Allora posso dire che ho una febbre moderata (**nuova conoscenza**)

L'Informatica

Nascita della Informatica

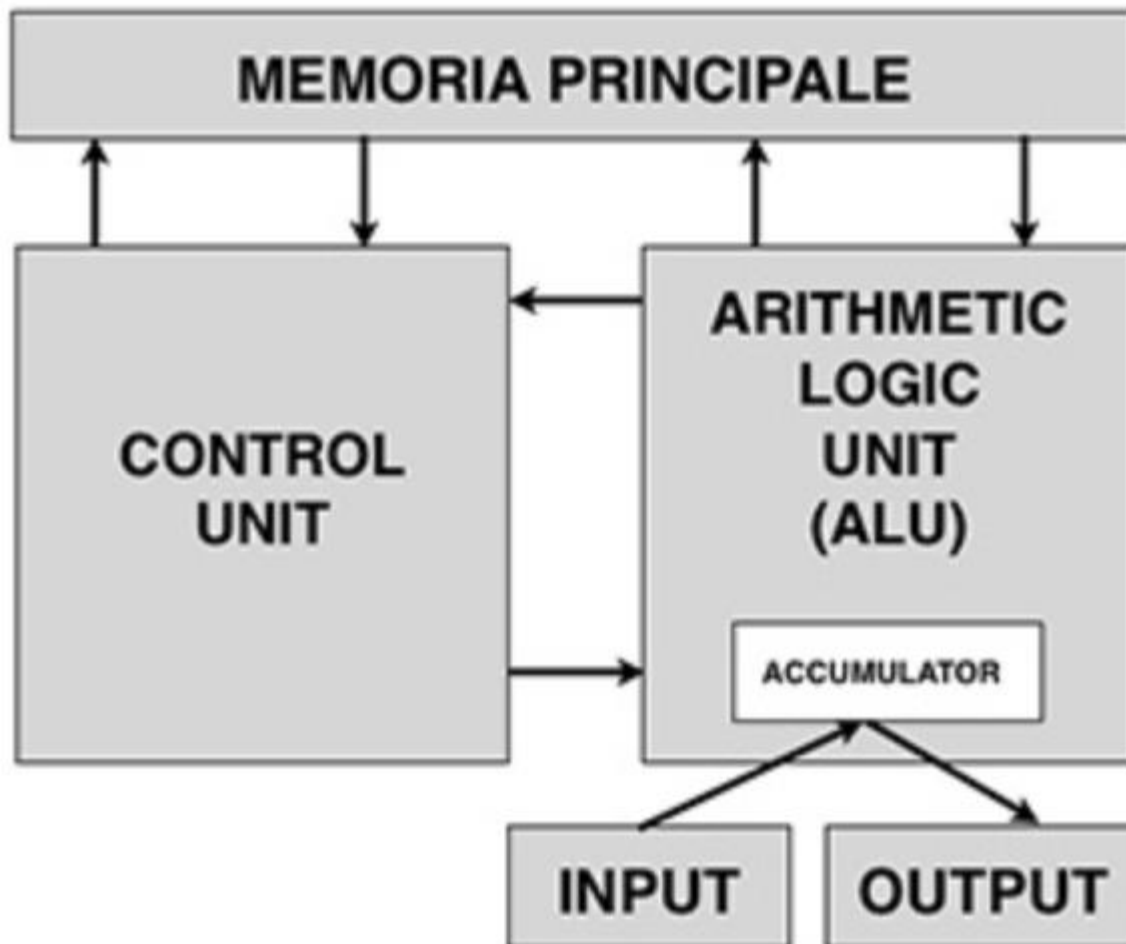


Macchina di Turing e concetto di computabilità

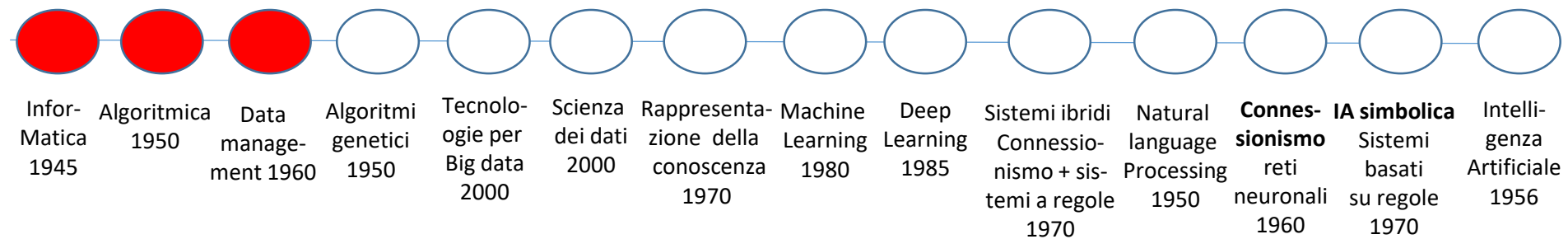


Problema della terminazione
Non tutto è calcolabile!

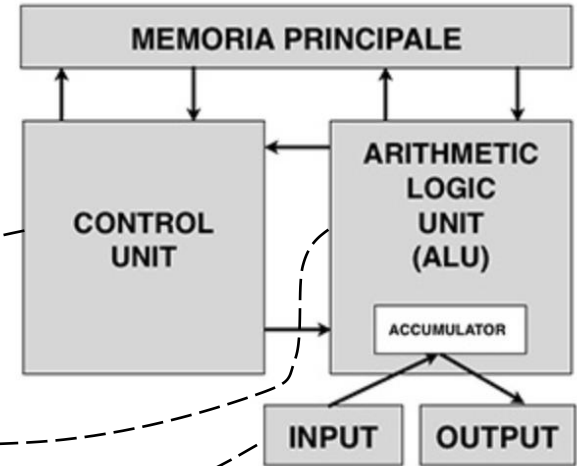
Macchina di von Neumann



Le due anime della Informatica: algoritmi e dati



A red triangle with vertices labeled A, B, and C. Vertex A is at the top left, vertex B is at the bottom left, and vertex C is at the top right. The triangle is equilateral.



Dati, tamburi, dischi

1932

L'austriaco G. Taushek,
sulla base dei principi
scoperti da Pleumer,
inventa il tamburo
magnetico.

La capacità di un tamburo
Lungo 20cm e di 10 cm di
diametro è di 500.000 bit.



Disco



I dati rappresentano la “realtà”

Per rappresentare la realtà serve un **modello**

Vogliamo rappresentare tre studenti, Carlo Batini, Ling Xu e John Smith, con i loro nomi e cognomi e la matricola, i corsi universitari di Analisi, Algoritmi e Logica, con gli anni in cui sono insegnati, e gli esami sostenuti dagli studenti, Analisi sostenuta da Batini con voto 25, Logica da Xu con voto 28 e ancora Analisi sostenuta da Smith con voto 30



Puntatore

Studente

Matricola	Nome	Cognome
13242	Carlo	Batini
24195	Ling	Xu
32845	John	Smith

Esame

Corso	Studente	Voto
		25
		28
		30

Corso

Nome Corso	Anno
Analisi	1
Algoritmi	1
Logica	2

I Dati, la Cenerentola dell'Informatica



Studente

Matricola	Nome	Cognome
13242	Carlo	Batini
24195	Ling	Xu
32845	John	Smith

Esame

Corso	Studente	Voto
		25
		28
		30

Corso

Nome Corso	Anno
Analisi	1
Algoritmi	1
Logica	2

I **dati digitali** come rappresentazione della realtà

Strutture matematiche,
linguistiche
o basate sui sensi



Processo di rappresentazione



Ho visto un bellissimo
tramonto in montagna,
era una giornata con
poche nuvole....



1970 – La svolta

A Relational Model of Data for Large Shared Data Banks

E. F. Codd
IBM Research Laboratory, San Jose, California

Future users of large data banks must be protected from having to know how the data is organized in the machine (the internal representation). A prompting service which supplies such information is not a satisfactory solution. Activities of users at terminals and most application programs should remain unaffected when the internal representation of data is changed and even when some aspects of the external representation are changed. Changes in data representation will often be needed as a result of changes in query, update, and report traffic and natural growth in the types of stored information.

Existing noninferential, formatted data systems provide users with tree-structured files or slightly more general network models of the data. In Section 1, inadequacies of these models are discussed. A model based on n -ary relations, a normal form for data base relations, and the concept of a universal data sublanguage are introduced. In Section 2, certain operations on relations (other than logical inference) are discussed and applied to the problems of redundancy and consistency in the user's model.

KEY WORDS AND PHRASES: data bank, data base, data structure, data organization, hierarchies of data, networks of data, relations, derivability, redundancy, consistency, composition, join, retrieval language, predicate calculus, security, data integrity

CR CATEGORIES: 3.70, 3.73, 3.75, 4.20, 4.22, 4.29

1. Relational Model and Normal Form

1.1. INTRODUCTION

This paper is concerned with the application of elementary relation theory to systems which provide shared access to large banks of formatted data. Except for a paper by Childs [1], the principal application of relations to data systems has been to deductive question-answering systems. Levein and Maron [2] provide numerous references to work in this area.

In contrast, the problems treated here are those of *data independence*—the independence of application programs and terminal activities from growth in data types and changes in data representation—and certain kinds of *data inconsistency* which are expected to become troublesome even in nondeductive systems.

The relational view (or model) of data described in Section 1 appears to be superior in several respects to the graph or network model [3, 4] presently in vogue for noninferential systems. It provides a means of describing data with its natural structure only—that is, without superimposing any additional structure for machine representation purposes. Accordingly, it provides a basis for a high level data language which will yield maximal independence between programs on the one hand and machine representation and organization of data on the other.

A further advantage of the relational view is that it forms a sound basis for treating derivability, redundancy, and consistency of relations—these are discussed in Section 2. The network model, on the other hand, has spawned a number of confusions, not the least of which is mistaking the derivation of connections for the derivation of relations (see remarks in Section 2 on the “connection trap”).

Finally, the relational view permits a clearer evaluation of the scope and logical limitations of present formatted data systems, and also the relative merits (from a logical standpoint) of competing representations of data within a single system. Examples of this clearer perspective are cited in various parts of this paper. Implementations of systems to support the relational model are not discussed.

1.2. DATA DEPENDENCIES IN PRESENT SYSTEMS

The provision of data description tables in recently developed information systems represents a major advance toward the goal of data independence [5, 6, 7]. Such tables facilitate changing certain characteristics of the data representation stored in a data bank. However, the variety of data representation characteristics which can be changed *without logically impairing some application programs* is still quite limited. Further, the model of data with which users interact is still cluttered with representational properties, particularly in regard to the representation of collections of data (as opposed to individual items). Three of the principal kinds of data dependencies which still need to be removed are: ordering dependence, indexing dependence, and access path dependence. In some systems these dependencies are not clearly separable from one another.

1.2.1. *Ordering Dependence.* Elements of data in a data bank may be stored in a variety of ways, some involving no concern for ordering, some permitting each element to participate in one ordering only, others permitting each element to participate in several orderings. Let us consider those existing systems which either require or permit data elements to be stored in at least one total ordering which is closely associated with the hardware-determined ordering of addresses. For example, the records of a file concerning parts might be stored in ascending order by part serial number. Such systems normally permit application programs to assume that the order of presentation of records from such a file is identical to (or is a subordering of) the

Il **modello relazionale**, un modello di spartana semplicità

Studente

Matricola	Nome	Cognome
13242	Carlo	Batini
24195	Ling	Xu
32845	John	Smith

Esame

Matricola	Codice Corso	Voto
13242	27	25
24195	77	28
32845	27	30

Corso

Codice	Nome Corso	Anno
27	Analisi	1
49	Algoritmi	1
77	Logica	2



Studente

Matricola	Nome	Cognome
13242	Carlo	Batini
24195	Ling	Xu
32845	John	Smith

Esame

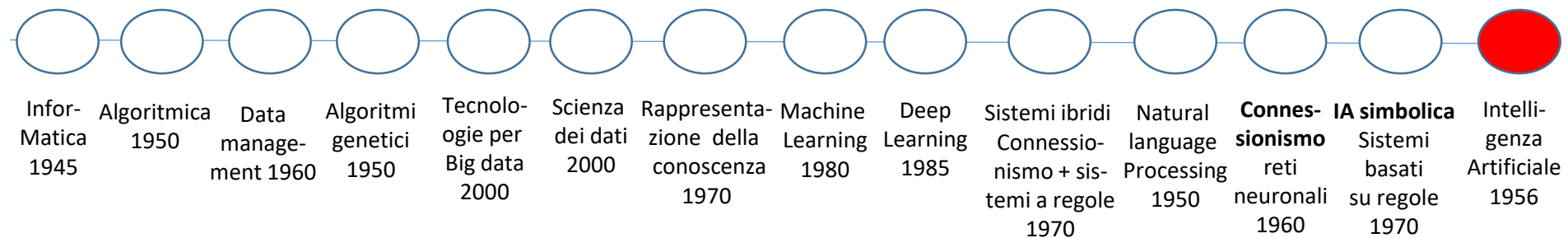
Corso	Studente	Voto
		25
		28
		30

Corso

Nome Corso	Anno
Analisi	1
Algoritmi	1
Logica	2

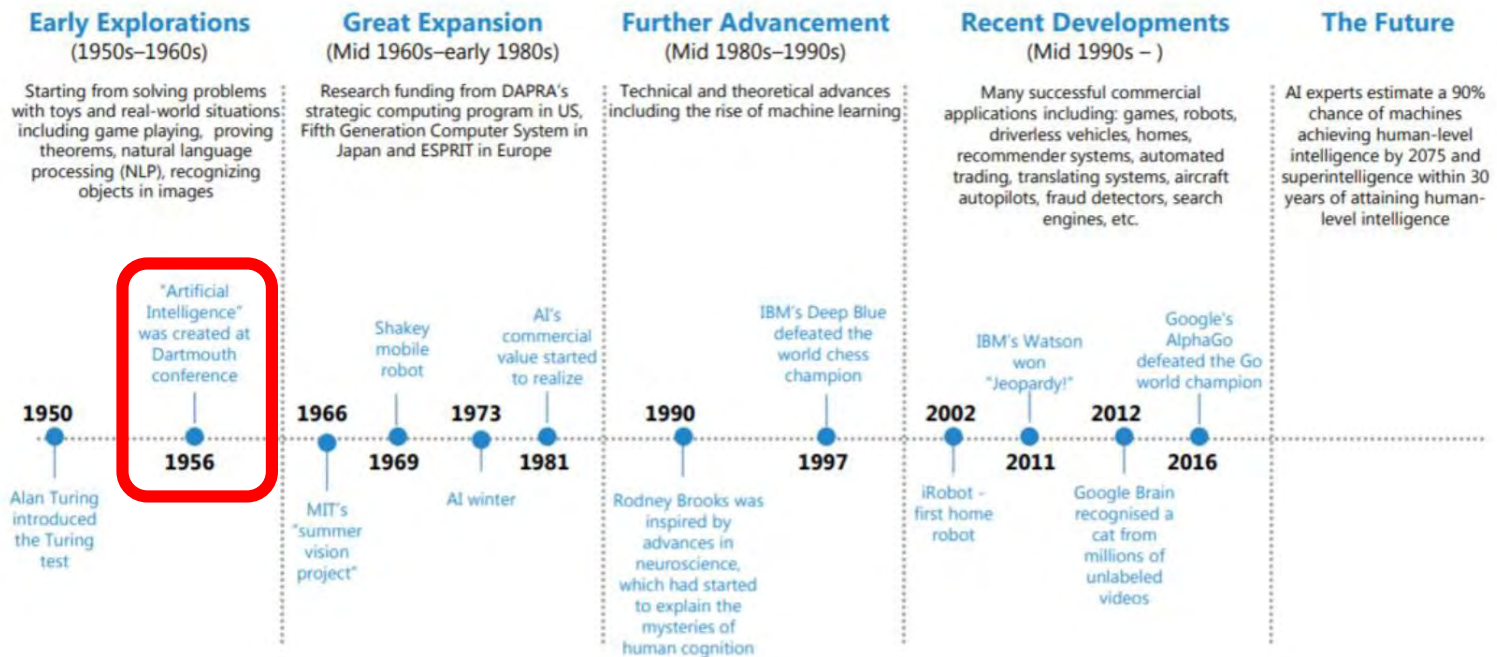
L' Intelligenza Artificiale i primi passi

La nascita della Intelligenza Artificiale (IA)



La IA nasce nel 1956

AI Evolution Over the Years



<https://www.slideshare.net/yanaiairona/vertex-perspectives-artificial-intelligence>

Una prima definizione di Intelligenza Artificiale

L'insieme di modelli, sistemi formali, tecniche, applicazioni che hanno lo scopo di

- **Riprodurre** nelle macchine digitali **l'intelligenza umana**, ovvero
- Far eseguire alle macchine digitali **compiti tipicamente svolti da umani**

Il gioco dell'imitazione

Mi propongo di considerare la questione: «Possono pensare le macchine?». Si dovrebbe cominciare col definire il significato dei termini macchina e pensare [...]. Invece di tentare una definizione [...] sostituirò la domanda con un'altra, che è strettamente analoga e che è espressa in termini non troppo ambigui.

Ciò può essere descritto nei termini di un gioco, che chiameremo «il gioco dell'imitazione». Questo viene giocato da tre persone, un uomo (A), una donna (B) e l'interrogante (C), che può essere dell'uno o dell'altro sesso.

Scopo del gioco per l'interrogante è quello di determinare quale delle altre due persone sia l'uomo e quale la donna. Egli le conosce con le etichette X e Y, ed alla fine del gioco darà la soluzione «X è A e Y è B» o la soluzione «X è B e Y è A». L'interrogante può far domande di questo tipo ad A e B:

C: «Vuol dirmi X, per favore, la lunghezza dei propri capelli?»

Ora supponiamo che X sia in effetti A, quindi A deve rispondere. Scopo di A nel gioco è quello di ingannare C e far sì che fornisca una identificazione errata. La sua risposta potrebbe perciò essere:

«I miei capelli sono tagliati "a la garçonne" ed i più lunghi sono di circa venticinque centimetri».

Le risposte, in modo che il tono di voce non possa aiutare l'interrogante, dovrebbero essere scritte, o, meglio ancora, battute a macchina. La soluzione migliore sarebbe quella di avere una telescrivente che mettesse in comunicazione le due stanze. Oppure le domande e risposte potrebbero essere ripetute da un intermediario.

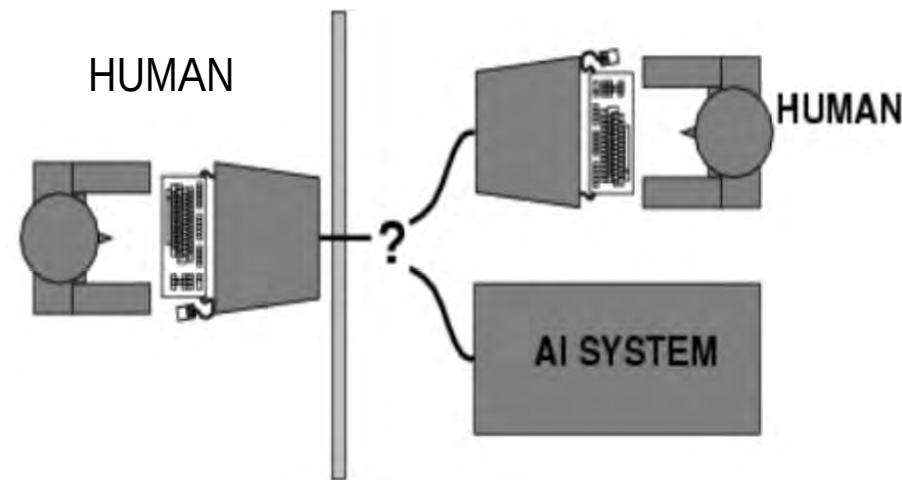
Scopo del gioco, per il terzo giocatore (B), è quello di aiutare l'interrogante. La migliore strategia per lei è quella di dare risposte veritiere. Essa può anche aggiungere alle sue risposte frasi come «Sono io la donna, non dargli ascolto», ma ciò non approderà a nulla dato che anche l'uomo può fare affermazioni analoghe.

Poniamo ora la domanda «Che cosa accadrà se una macchina prenderà il posto di A nel gioco?». L'interrogante darà una risposta errata altrettanto spesso di quando il gioco viene giocato tra un uomo e una donna?

Queste domande sostituiscono quella originale: «Possono pensare le macchine?» [...].

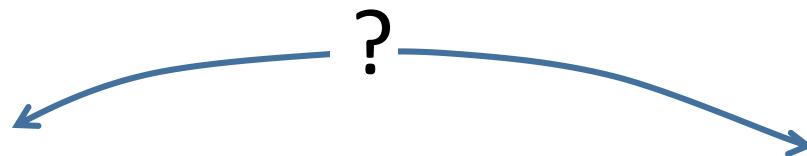
«Credo che entro circa 50 anni sarà possibile programmare calcolatori con una capacità di memorizzazione di circa 10^9 , per fare giocare loro il gioco dell'imitazione così bene che un esaminatore non avrà più del 70% di probabilità di compiere l'identificazione esatta dopo 5 minuti di interrogazione».

Il test di Turing



Una variante del Test di Turing

La frase da “Le città invisibili” di Calvino tradotta da un umano e da una applicazione di IA

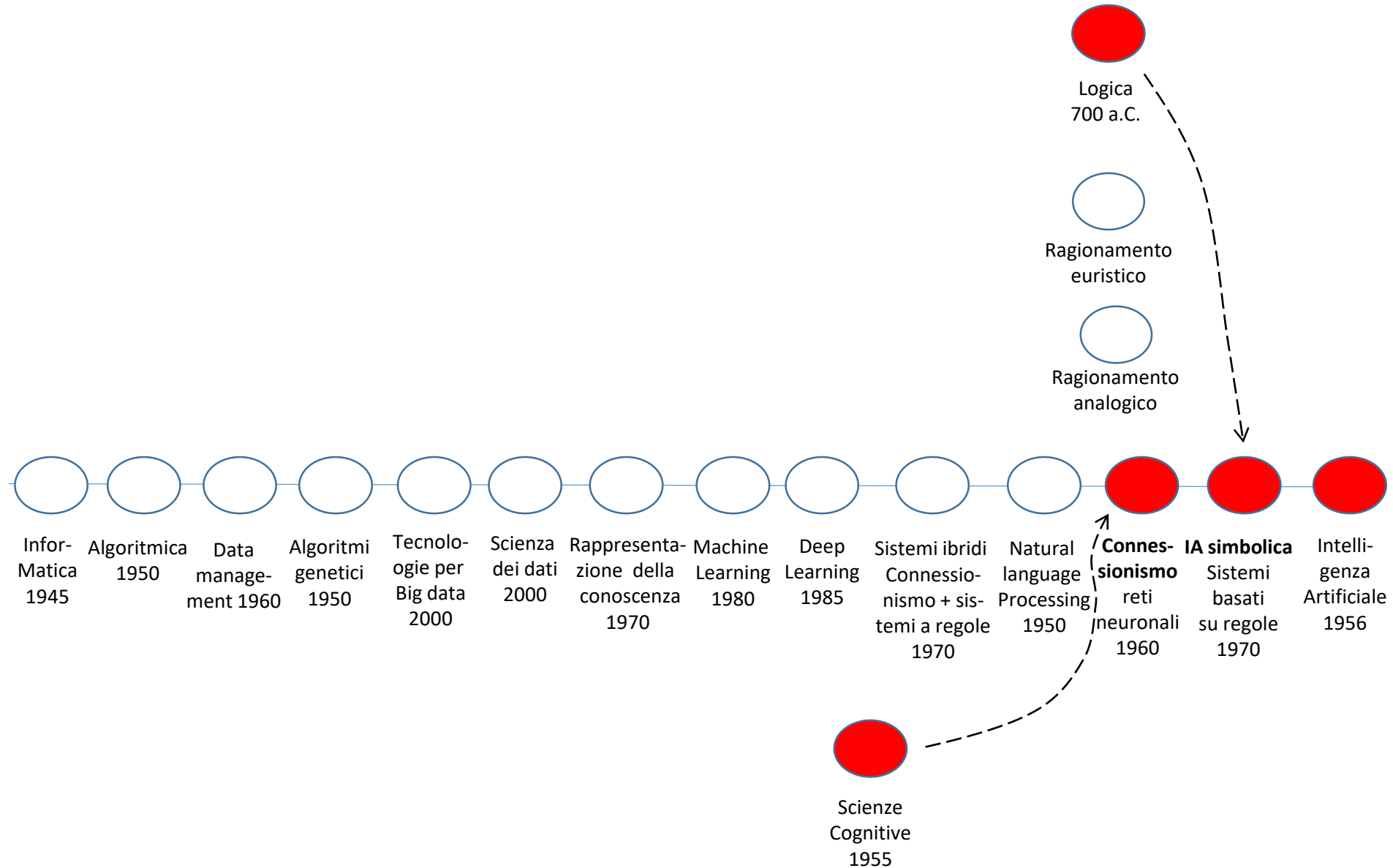


“Arrivando a ogni nuova città il viaggiatore ritrova un suo passato che non sapeva più d'avere: l'estraneità di ciò che non sei più o non possiedi più t'aspetta al varco nei luoghi estranei e non posseduti.”

“By reaching every new city, the traveler finds a past that no longer knew: The disenchantment of what you are no longer, or you no longer have, awaits the crossing in foreign and non-owned places.”

“Arriving at each new city, the traveler finds again a past of his that he did not know he had: the foreignness of what you no longer are or no longer possess lies in wait for you in foreign, unpossessed places.”

Le due scuole della prima Intelligenza Artificiale



IA simbolica (sistemi a regole)

- Usualmente, quando esco, se fuori piove, porto l'ombrello
- Oggi devo uscire e fuori piove $\rightarrow$ Porto l'ombrello o no?

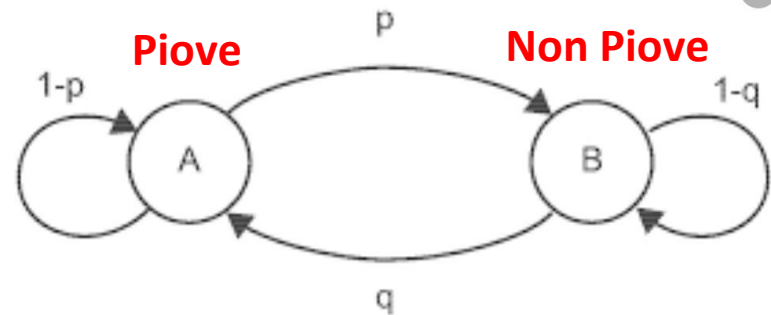
Modus Ponens nel Calcolo dei Predicati

Piove (t) AND Esco (t) $\rightarrow$ Porto l'ombrello (t)

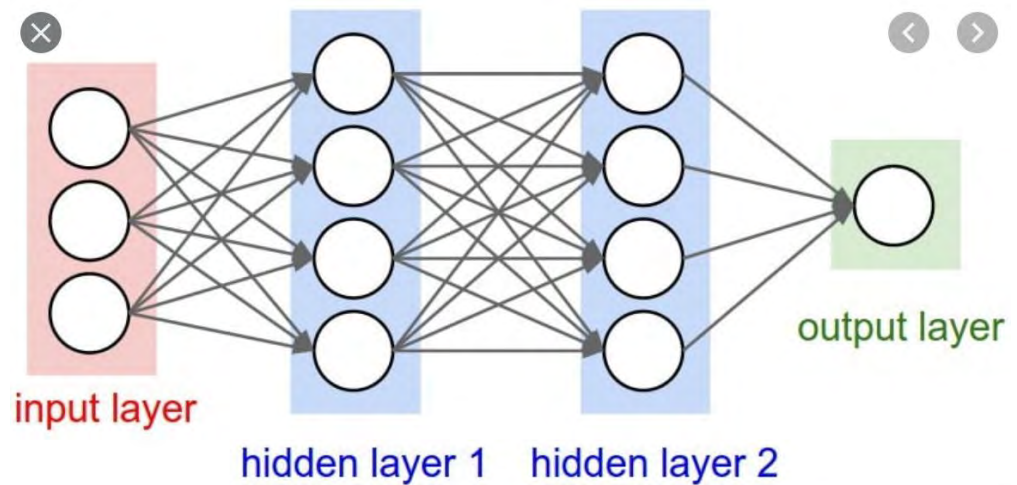
> Piove (oggi) AND Esco (oggi)

> Sì, Porto l'ombrello (oggi)

IA connessionistica



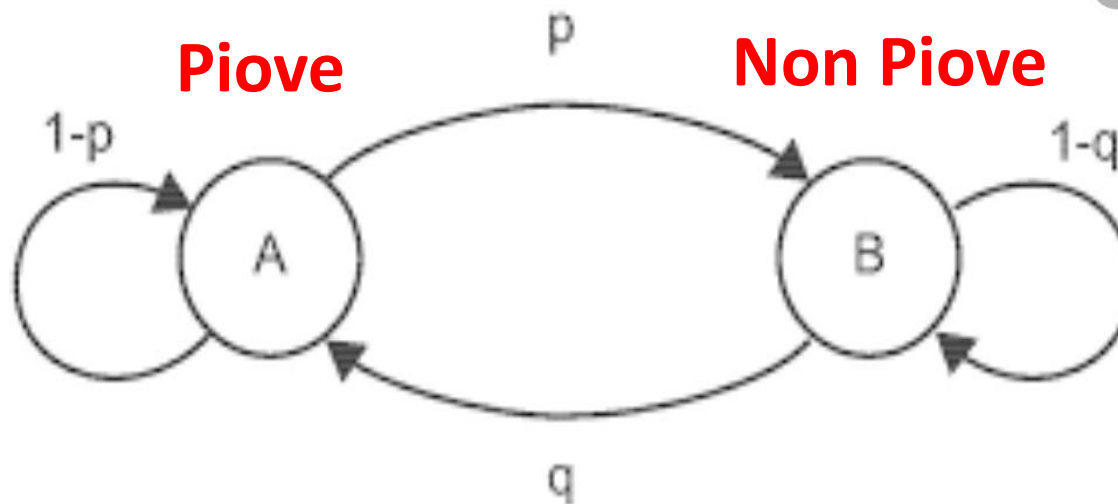
Rete Markoviana



Rete Neurale Artificiale

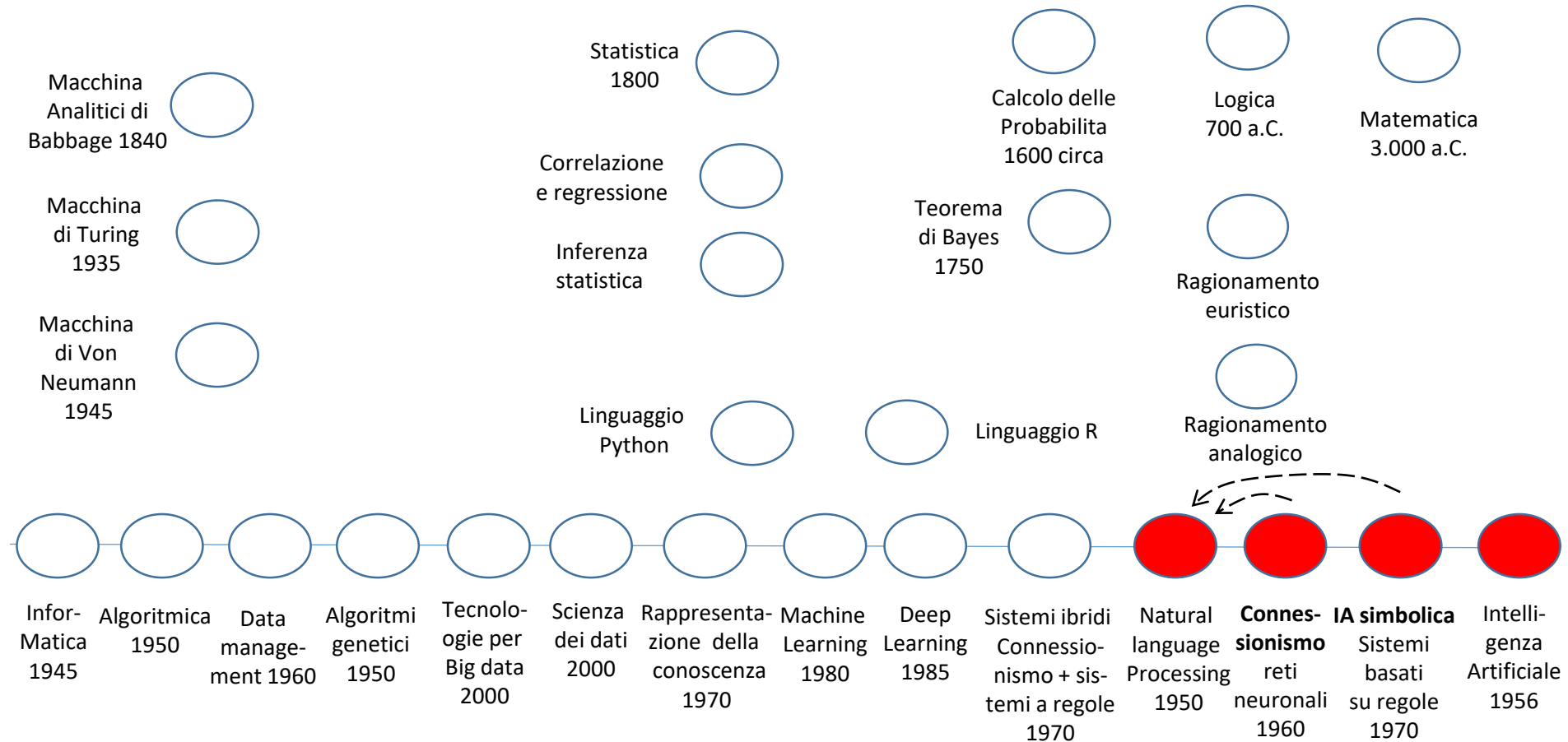
Il teorema di Bayes, ovvero come convivere con la incertezza

$$P(\text{evento}/\text{dati}) \propto P(\text{dati}/\text{evento}) \times P(\text{evento})$$



Rete Bayesiana

Il grande sforzo degli anni 60/70



L'elaborazione del linguaggio naturale

Grammatiche di Chomsky

- $\langle \text{numero intero} \rangle := \langle \text{cifra} \rangle / \langle \text{cifra} \rangle \langle \text{numero intero} \rangle$
- $\langle \text{cifra} \rangle := 1/2/3/4/5/6/7/8/9/10$

Tecniche di traduzione

es. Dall'italiano all'inglese

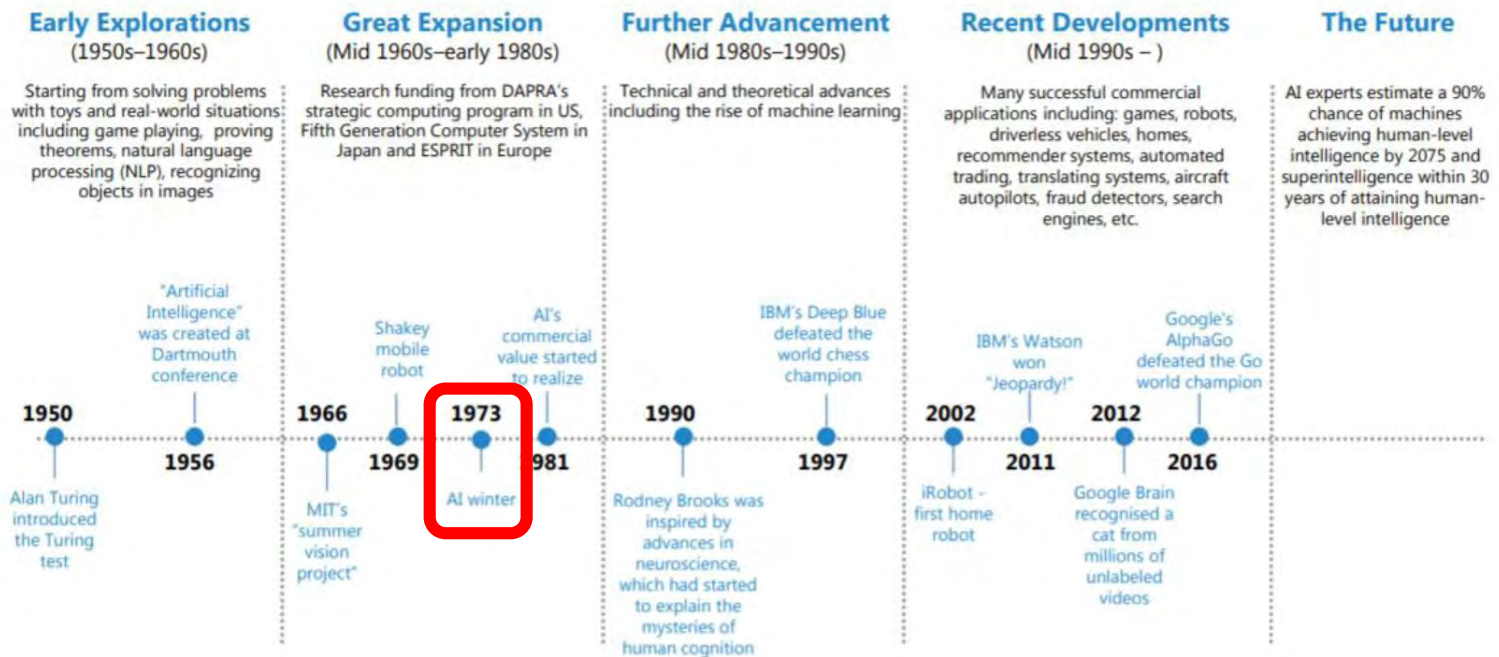
Input – Garibaldi nacque a Nizza?

1. Isola le single parole, individuando il lessico
2. Risolvi eventuali sinonimi e omonimi
3. Individua la sintassi della frase
4. Trasforma la espressione dalla sintassi italiana alla sintassi inglese (es. Il verbo in italiano è nella seconda posizione, in inglese nella prima)
5. Traduci le parole nel lessico inglese

Output – Was Garibaldi born in Nice?

L'inverno della IA

AI Evolution Over the Years



<https://www.slideshare.net/yanairona/vertex-perspectives-artificial-intelligence>

L'inverno della IA - 1

- In 1969, Minsky and Papert's book *Perceptrons* was published. **It was a harsh critique on Rosenblatt's perceptrons.** Minsky and Papert proved that **perceptrons could only be trained to solve linear separable problems**

L'inverno della IA - 2

- The Lighthill report (published in 1973) was an evaluation of the current state of AI at that time written for the British Science Research Council.
- The report came to the conclusion that the promises of AI researchers were exaggerated: **“in no part of the field have discoveries made so far produced the major impact that was then promised.”**
- Though it pointed out that **the most disappointing area of research had been machine translation**, “... where enormous sums have been spent with very little useful results...”
- James Lighthill, the author of the report, thought that **the failure to defeat the “combinatorial explosion” was at the heart of the issue.**

L'intelligenza umana è sfaccettata e poliedrica

L'intelligenza emotiva è un aspetto dell'intelligenza legato alla capacità di riconoscere, utilizzare, comprendere e gestire in modo consapevole le proprie ed altrui emozioni

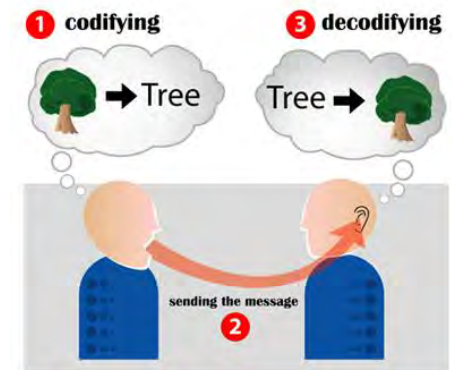


Gli umani acquisiscono la conoscenza da una pluralità di fonti

L'esperienza



La condivisione



L'inferenza

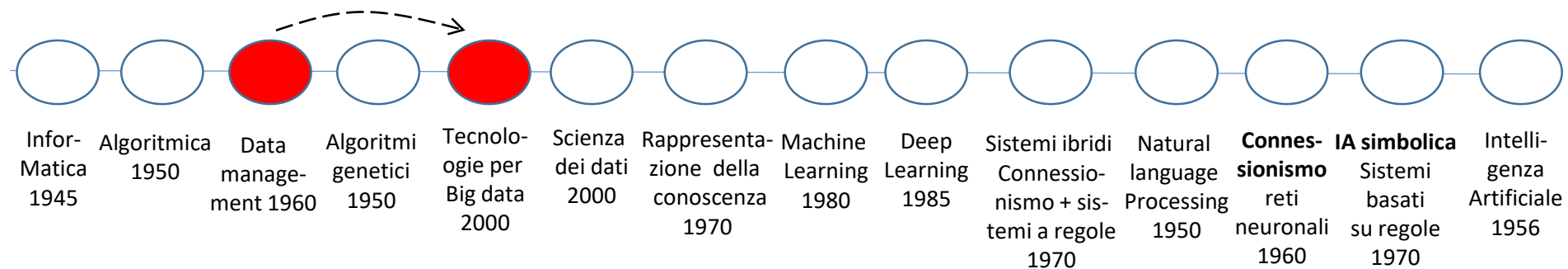


Le astrazioni nella vita reale



- Guarda quella cosa
- Guarda quell'albero
- Guarda quel pino
- Guarda quel pino romano

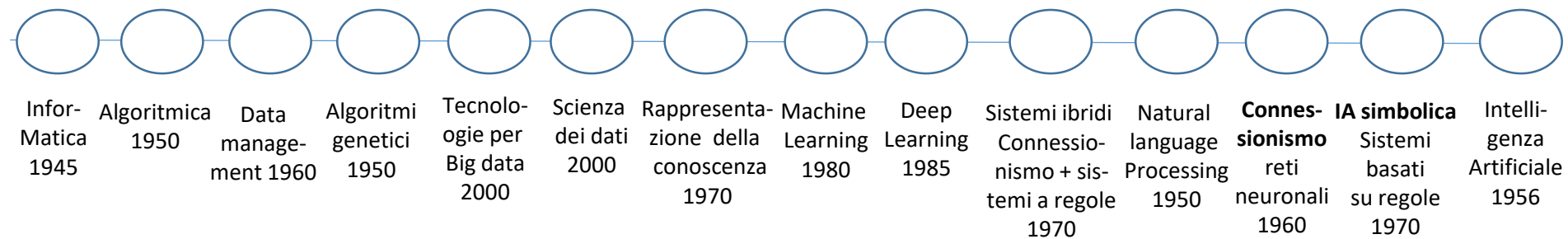
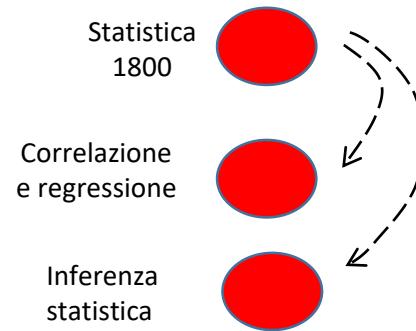
Torniamo ai dati: dai piccoli dati ai grandi dati



Pochi, pochissimi dati



I primi vagiti della Statistica



I Geografi Babilonesi e la previsione delle eclissi



Le prime applicazioni della Statistica

The Year of the Issue		1961	1962	1963	1964	1965	1966	1967	1968	1969	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025	2026	2027	2028	2029	2030	2031	2032	2033	2034	2035	2036	2037	2038	2039	2040	2041	2042	2043	2044	2045	2046	2047	2048	2049	2050	2051	2052	2053	2054	2055	2056	2057	2058	2059	2060	2061	2062	2063	2064	2065	2066	2067	2068	2069	2070	2071	2072	2073	2074	2075	2076	2077	2078	2079	2080	2081	2082	2083	2084	2085	2086	2087	2088	2089	2090	2091	2092	2093	2094	2095	2096	2097	2098	2099	2100	2101	2102	2103	2104	2105	2106	2107	2108	2109	2110	2111	2112	2113	2114	2115	2116	2117	2118	2119	2120	2121	2122	2123	2124	2125	2126	2127	2128	2129	2130	2131	2132	2133	2134	2135	2136	2137	2138	2139	2140	2141	2142	2143	2144	2145	2146	2147	2148	2149	2150	2151	2152	2153	2154	2155	2156	2157	2158	2159	2160	2161	2162	2163	2164	2165	2166	2167	2168	2169	2170	2171	2172	2173	2174	2175	2176	2177	2178	2179	2180	2181	2182	2183	2184	2185	2186	2187	2188	2189	2190	2191	2192	2193	2194	2195	2196	2197	2198	2199	2200	2201	2202	2203	2204	2205	2206	2207	2208	2209	2210	2211	2212	2213	2214	2215	2216	2217	2218	2219	2220	2221	2222	2223	2224	2225	2226	2227	2228	2229	2230	2231	2232	2233	2234	2235	2236	2237	2238	2239	2240	2241	2242	2243	2244	2245	2246	2247	2248	2249	2250	2251	2252	2253	2254	2255	2256	2257	2258	2259	2260	2261	2262	2263	2264	2265	2266	2267	2268	2269	2270	2271	2272	2273	2274	2275	2276	2277	2278	2279	2280	2281	2282	2283	2284	2285	2286	2287	2288	2289	2290	2291	2292	2293	2294	2295	2296	2297	2298	2299	2300	2301	2302	2303	2304	2305	2306	2307	2308	2309	2310	2311	2312	2313	2314	2315	2316	2317	2318	2319	2320	2321	2322	2323	2324	2325	2326	2327	2328	2329	2330	2331	2332	2333	2334	2335	2336	2337	2338	2339	2340	2341	2342	2343	2344	2345	2346	2347	2348	2349	2350	2351	2352	2353	2354	2355	2356	2357	2358	2359	2360	2361	2362	2363	2364	2365	2366	2367	2368	2369	2370	2371	2372	2373	2374	2375	2376	2377	2378	2379	2380	2381	2382	2383	2384	2385	2386	2387	2388	2389	2390	2391	2392	2393	2394	2395	2396	2397	2398	2399	2400	2401	2402	2403	2404	2405	2406	2407	2408	2409	2410	2411	2412	2413	2414	2415	2416	2417	2418	2419	2420	2421	2422	2423	2424	2425	2426	2427	2428	2429	2430	2431	2432	2433	2434	2435	2436	2437	2438	2439	2440	2441	2442	2443	2444	2445	2446	2447	2448	2449	2450	2451	2452	2453	2454	2455	2456	2457	2458	2459	2460	2461	2462	2463	2464	2465	2466	2467	2468	2469	2470	2471	2472	2473	2474	2475	2476	2477	2478	2479	2480	2481	2482	2483	2484	2485	2486	2487	2488	2489	2490	2491	2492	2493	2494	2495	2496	2497	2498	2499	2500	2501	2502	2503	2504	2505	2506	2507	2508	2509	2510	2511	2512	2513	2514	2515	2516	2517	2518	2519	2520	2521	2522	2523	2524	2525	2526	2527	2528	2529	2530	2531	2532	2533	2534	2535	2536	2537	2538	2539	2540	2541	2542	2543	2544	2545	2546	2547	2548	2549	2550	2551	2552	2553	2554	2555	2556	2557	2558	2559	2560	2561	2562	2563	2564	2565	2566	2567	2568	2569	2570	2571	2572	2573	2574	2575	2576	2577	2578	2579	2580	2581	2582	2583	2584	2585	2586	2587	2588	2589	2590	2591	2592	2593	2594	2595	2596	2597	2598	2599	2600	2601	2602	2603	2604	2605	2606	2607	2608	2609	2610	2611	2612	2613	2614	2615	2616	2617	2618	2619	2620	2621	2622	2623	2624	2625	2626	2627	2628	2629	2630	2631	2632	2633	2634	2635	2636	2637	2638	2639	2640	2641	2642	2643	2644	2645	2646	2647	2648	2649	2650	2651	2652	2653	2654	2655	2656	2657	2658	2659	2660	2661	2662	2663	2664	2665	2666	2667	2668	2669	2670	2671	2672	2673	2674	2675	2676	2677	2678	2679	2680	2681	2682	2683	2684	2685	2686	2687	2688	2689	2690	2691	2692	2693	2694	2695	2696	2697	2698	2699	2700	2701	2702	2703	2704	2705	2706	2707	2708	2709	2710	2711	2712	2713	2714	2715	2716	2717	2718	2719	2720	2721	2722	2723	2724	2725	2726	2727	2728	2729	2730	2731	2732	2733	2734	2735	2736	2737	2738	2739	2740	2741	2742	2743	2744	2745	2746	2747	2748	2749	2750	2751	2752	2753	2754	2755	2756	2757	2758	2759	2760	2761	2762	2763	2764	2765	2766	2767	2768	2769	2770	2771	2772	2773	2774	2775	2776	2777	2778	2779	2780	2781	2782	2783	2784	2785	2786	2787	2788	2789	2790	2791	2792	2793	2794	2795	2796	2797	2798	2799	2800	2801	2802	2803	2804	2805	2806	2807	2808	2809	2810	2811	2812	2813	2814	2815	2816	2817	2818	2819	2820	2821	2822	2823	2824	2825	2826	2827	2828	2829	2830	2831	2832	2833	2834	2835	2836	2837	2838	2839	2840	2841	2842	2843	2844	2845	2846	2847	2848	2849	2850	2851	2852	2853	2854	2855	2856	2857	2858	2859	2860	2861	2862	2863	2864	2865	2866	2867	2868	2869	2870	2871	2872	2873	2874	2875	2876	2877	2878	2879	2880	2881	2882	2883	2884	2885	2886	2887	2888	2889	2890	2891	2892	2893	2894	2895	2896	2897	2898	2899	2900	2901	2902	2903	2904	2905	2906	2907	2908	2909	2910	2911	2912	2913	2914	2915	2916	2917	2918	2919	2920	2921	2922	2923	2924	2925	2926	2927	2928	2929	2930	2931	2932	2933	2934	2935	2936	2937	2938	2939	2940	2941	2942	2943	2944	2945	2946	2947	2948	2949	2950	2951	2952	2953	2954	2955	2956	2957	2958	2959	2960	2961	2962	2963	2964	2965	2966	2967	2968	2969	2970	2971	2972	2973	2974	2975	2976	2977	2978	2979	2980	2981	2982	2983	2984	2985	2986	2987	2988	2989	2990	2991	2992	2993	2994	2995	2996	2997	2998	2999	3000
Aluminum and Silica	131	149	171	191	209	224	234	239	240	239	234	224	209	189	169	149	129	109	89	69	49	29	9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																					

(119) Note, where there follows a second Number under any year, it denotes those which died that year of the Plague.

Anno Dom.	Wentimitor		Isington		Lambeth		Stepney		B.
	Bur.	Ch.	B.	C.	B.	Ch.	Bur.	Ch.	
1636	1107	556	99	56	213	137	1895	881	584
1637	442		30		45		909		242
1638	903	496	94	72	173	137	952	838	183
	301		17		18		153		16
1639	1021	563	110	49	221	140	1209	908	255
	126				8		11		
1640	546	543	88	5	195	132	970	956	187
	4		2		187		2		
1641	754	665	94	54	186	142	1106	983	189
	62		3		6		117		
1642	697	625	92	76	168	137	1250	1037	170
	40		98		9		70		
1643	671	630	96	71	149	124	1270	1158	160
	37		105		12		20		
1644	666	592	4	69	177	114	1167	1013	44
	25		1		45		83		
1645	570	429	61	55	115	105	1187	933	123
	35		8		8		269		44
1646	621	444	55	63	146	114	1171	873	183
	62		6		3		150		18
1647	691	503	84	61	137	108	1230	960	156
	76		8		5		97		14
1648	739	464	108	56	161	94	1126	926	120
	114		12		25		127		28
1649	561	384	68	46	87	57	837	767	
	41		4				31		
	558	333	90	44	131	55	838	625	

The Table by <i>Decads</i> of years for the Country Parish.							
Decads of Years	Married	Christened		Both	Buried		Both
		Males	Females		Males	Females	
15 { 69 78	190	312	302	614	214	221	435
15 { 79 88	185	328	309	637	287	302	589
15 { 89 98	175	342	274	616	337	284	621
1 { 599 608	181	366	377	743	249	219	468
16 { 09 18	197	417	358	775	338	386	724
16 { 19 28	168	368	373	741	305	306	611
16 { 29 38	153	418	413	831	317	319	636
16 { 39 48	137	351	357	708	375	383	758
16 { 49 58	182	354	320	674	218	220	438
	1598	3256	3083	6339	2640	2640	5280

males

6199

6041

6033

5730
6120

5566

35697

594

Il colera a Londra nel 1854 e le scoperte di John Snow: nasce la **correlazione**



Con pochi dati si fanno poche cose



Studio di caso 1 - Prevedere l'inquinamento



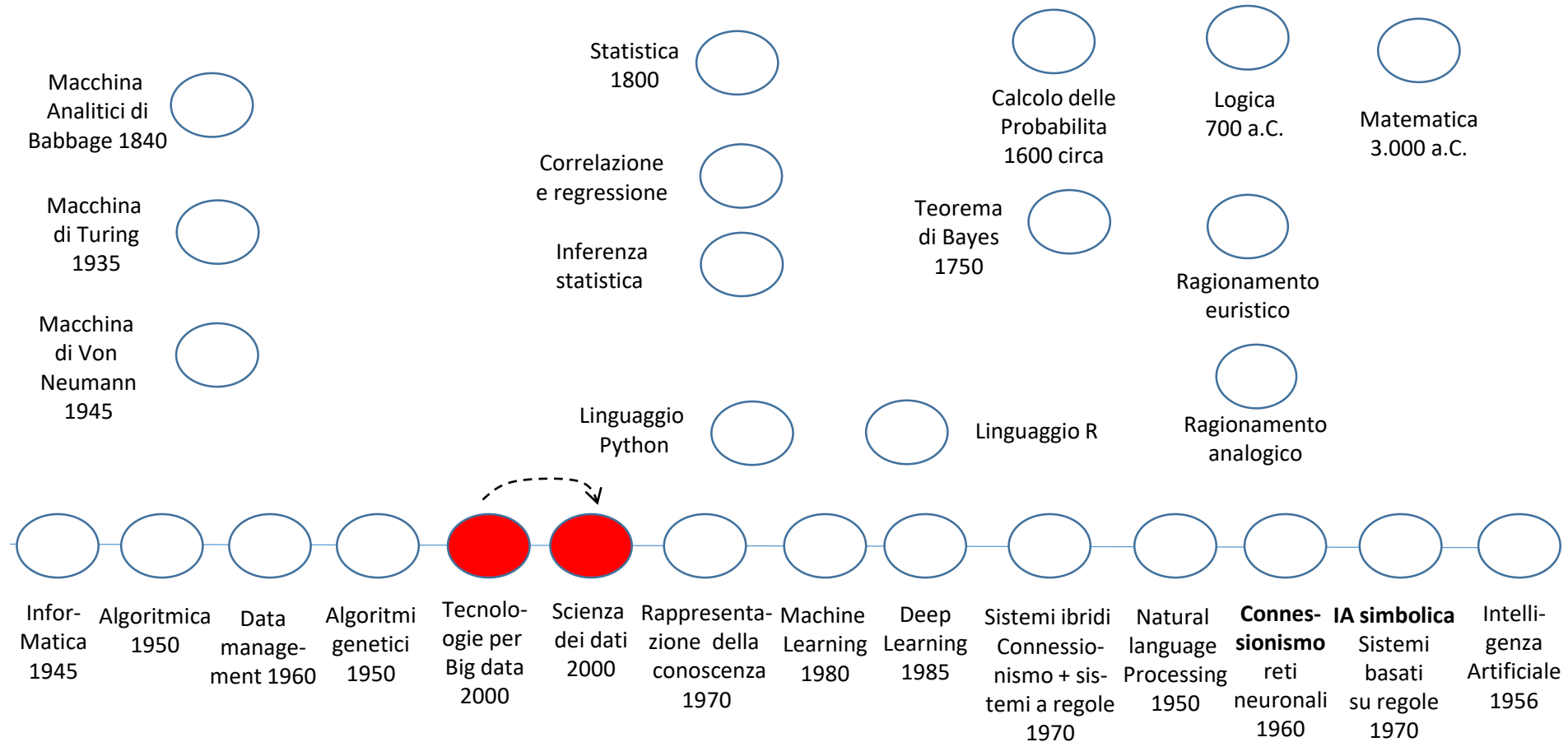
Prevedere l'inquinamento Breezometer



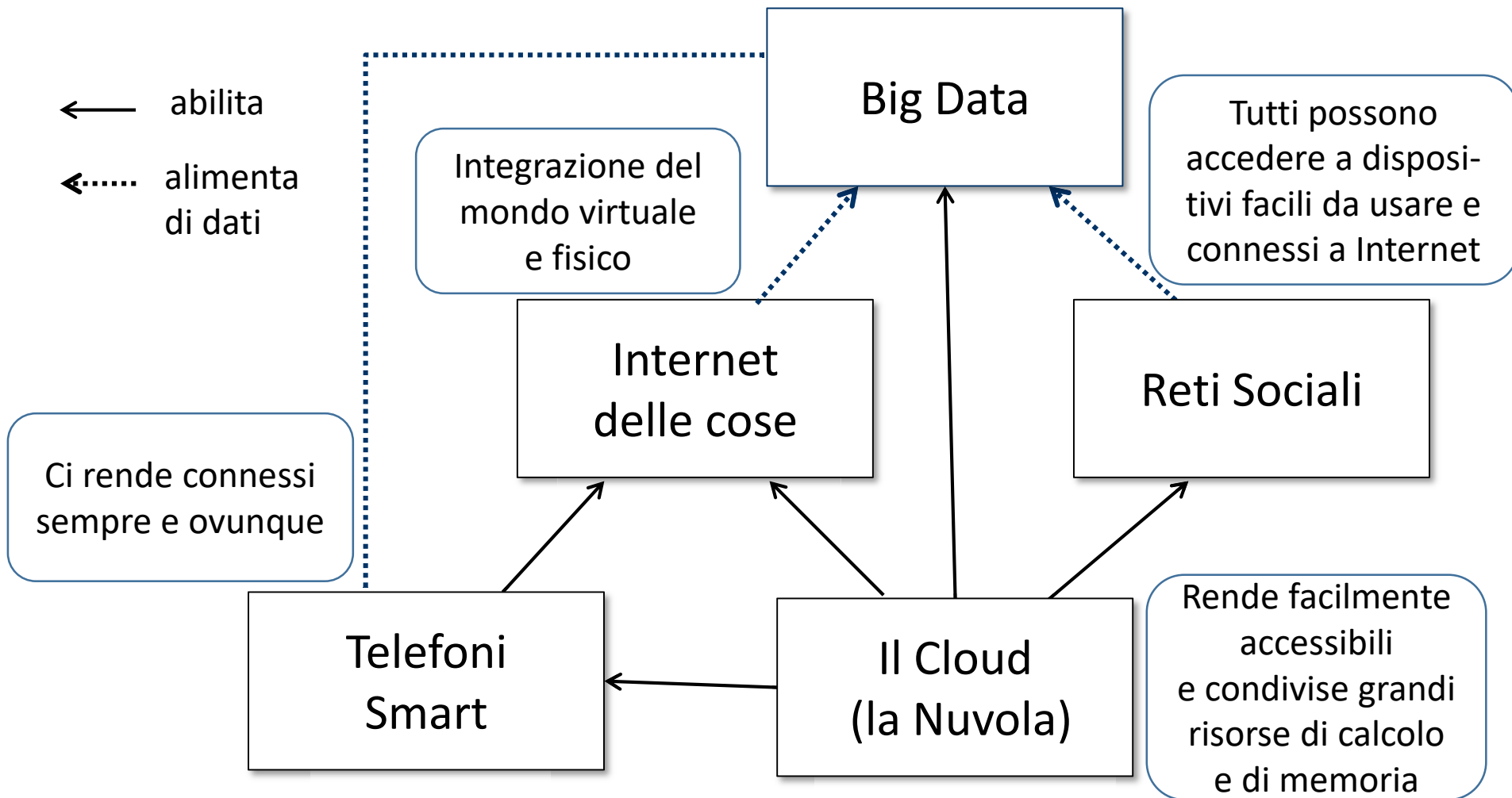
Breezometer



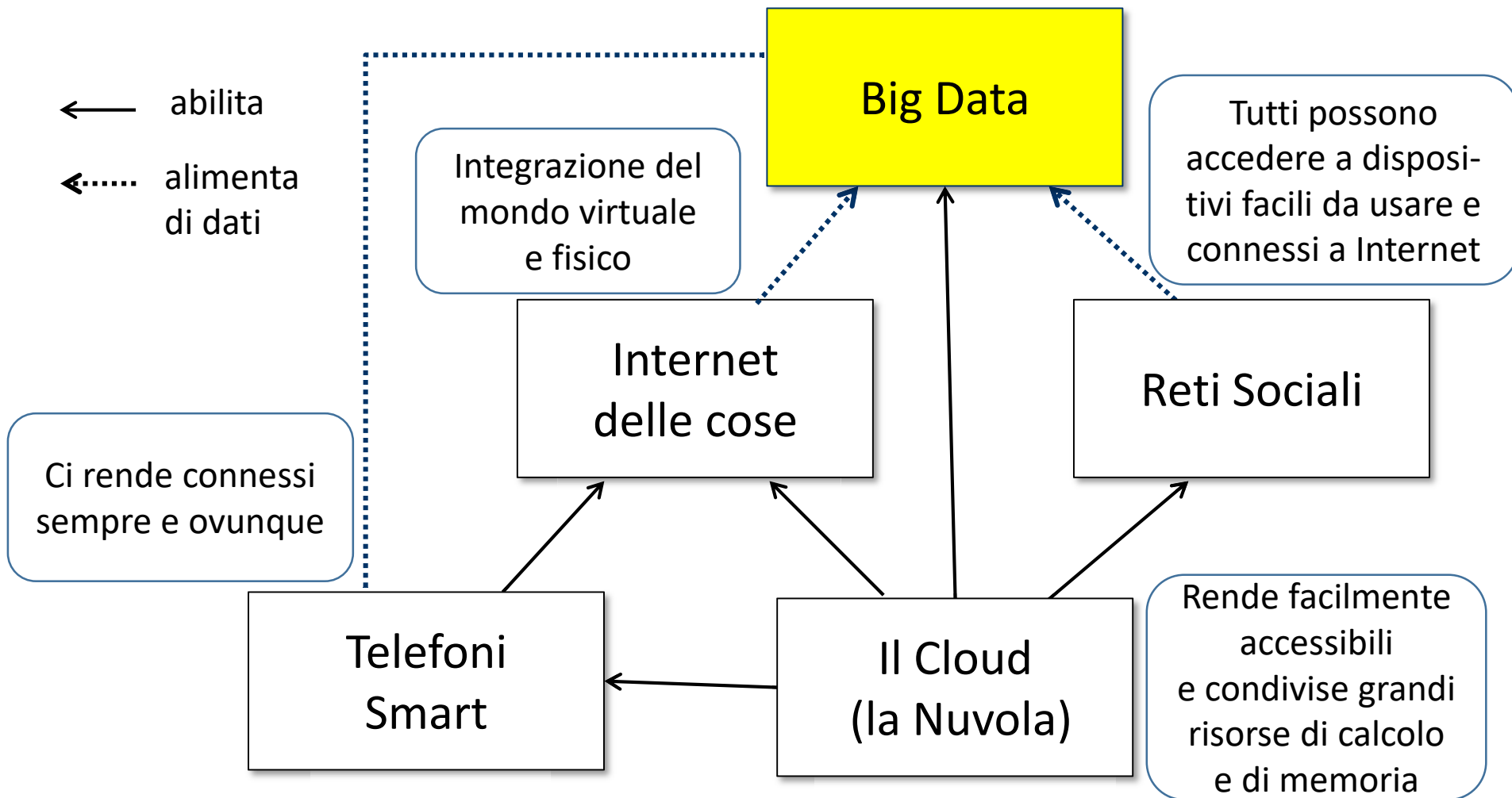
Le tecnologie che hanno dato impulso alla Scienza dei dati



Le cinque grandi tecnologie del nuovo mondo dei dati digitali



Le cinque grandi tecnologie del nuovo mondo dei dati digitali



Principali caratteristiche dei Big Data

- **Volume**, la quantità di dati che possono essere raccolti, memorizzati, elaborati
- **Velocità**, la velocità con cui i dati cambiano
- **Varietà**, la diversificazione del formato dei dati, da strutturati a non strutturati (testi) a immagini
- **Veridicità**, quanto i dati aderiscono accuratamente al fenomeno che descrivono

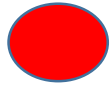
Principali caratteristiche dei Big Data

- **Volume**, la quantità di dati che possono essere raccolti, memorizzati, elaborati
- **Velocità**, la velocità con cui i dati cambiano
- **Varietà**, la diversificazione del formato dei dati, da strutturati a non strutturati (testi) a immagini
- **Veridicità**, quanto i dati aderiscono accuratamente al fenomeno che descrivono

PB = 10^7 / EB = 10^{18} byte

I quattro domini dei Big data nel 2025

Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction Real-time processing Massive volumes	Topic and sentiment mining Metadata analysis	Limited requirements	Heterogeneous data and analysis Variant calling, ~2 trillion central processing unit (CPU) hours All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement



Problema – Fissato il giorno di un viaggio, Trovare Il giorno in cui il biglietto costa meno

Milano

↔

Dar es Salaam DAR

📅 gio 17 gen

<

>

Scegli viaggio a Dar es Salaam

Riepilogo del viaggio

Bagagli ▾ Scali ▾ Compagnie aeree ▾ Prezzo ▾ Orari ▾ Aeroporti di scalo ▾ Altri ▾

Suggerimenti sui voli

Date
Guarda i prezzi dei voli in date simili

Grafico dei prezzi
Esplora le tendenze dei prezzi dei viaggi con destinazione Dar es Salaam

Aeroporti
Confronta i prezzi per gli aeroporti vicino a Dar es Salaam

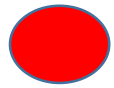
Suggerimenti
Volare in premium economy al costo di 987 €

Voli migliori ⓘ

Il prezzo totale include tasse e commissioni per 1 adulto. Potrebbero essere applicate [tariffe per bagagli aggiuntivi](#) e altre commissioni.

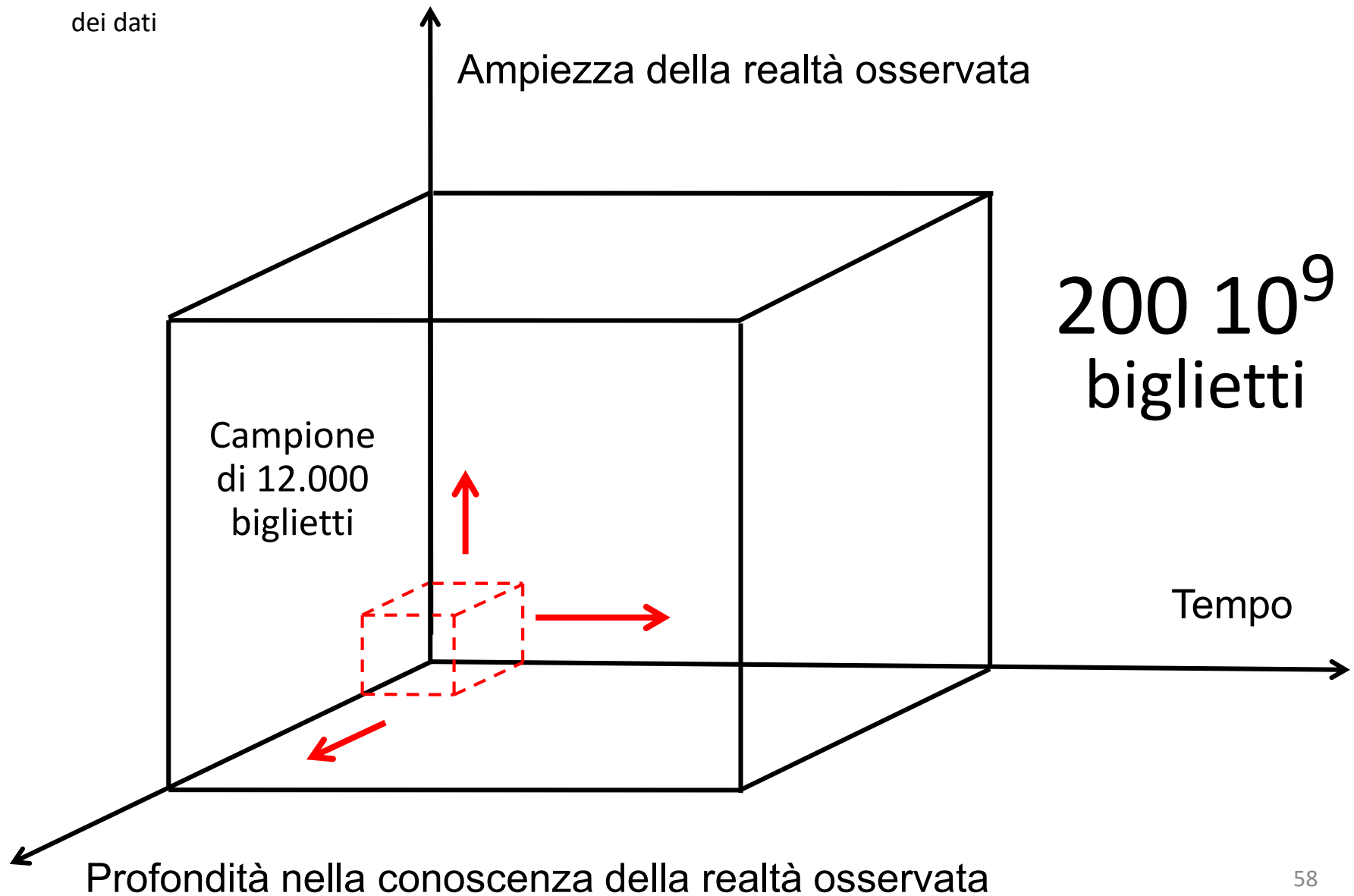
Ordina per: ↑↓

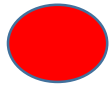
	11:00 - 08:05^{*1} Turkish Airlines	19 h 5 min MXP-DAR	2 scali ⚠ IST, LUN	348 €	▾
	18:55 - 03:05^{*2} Turkish Airlines	30 h 10 min MXP-DAR	1 scalo ⚠ 19 h 55 min IST	348 €	▾
	22:15 - 15:30^{*1} Qatar Airways	15 h 15 min MXP-DAR	1 scalo 2 h 55 min DOH	572 €	▾



Scienza
dei dati

Il modello predittivo di Oren Etzioni

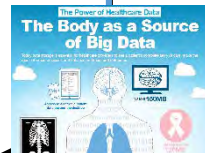
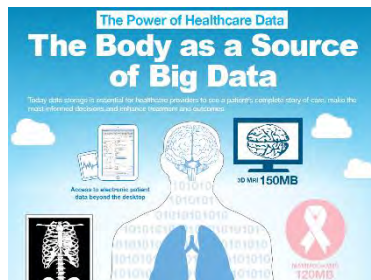
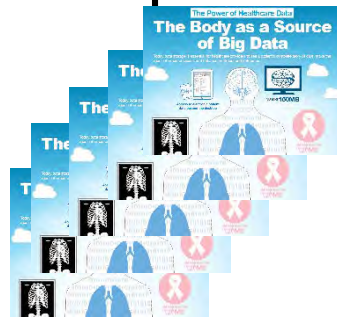




Scienza
dei dati

Il genoma umano

Ampiezza della realtà osservata



Tempo

Profondità nella conoscenza della realtà osservata

Principali caratteristiche dei Big Data

- **Volume**, la qualità di dati che possono essere raccolti, memorizzati, elaborati
- **Velocità**, la velocità con cui i dati cambiano
- **Varietà**, la diversificazione del formato dei dati, da strutturati a non strutturati (testi) a immagini
- **Veridicità**, quanto i dati aderiscono accuratamente al fenomeno che descrivono

In quanti modi si può rappresentare Portofino...

Luogo	Nazione	Popolazione	Attività economica
Portofino	Italia	7.000	Turismo

Tabella

Immagine



Mappa



Testo

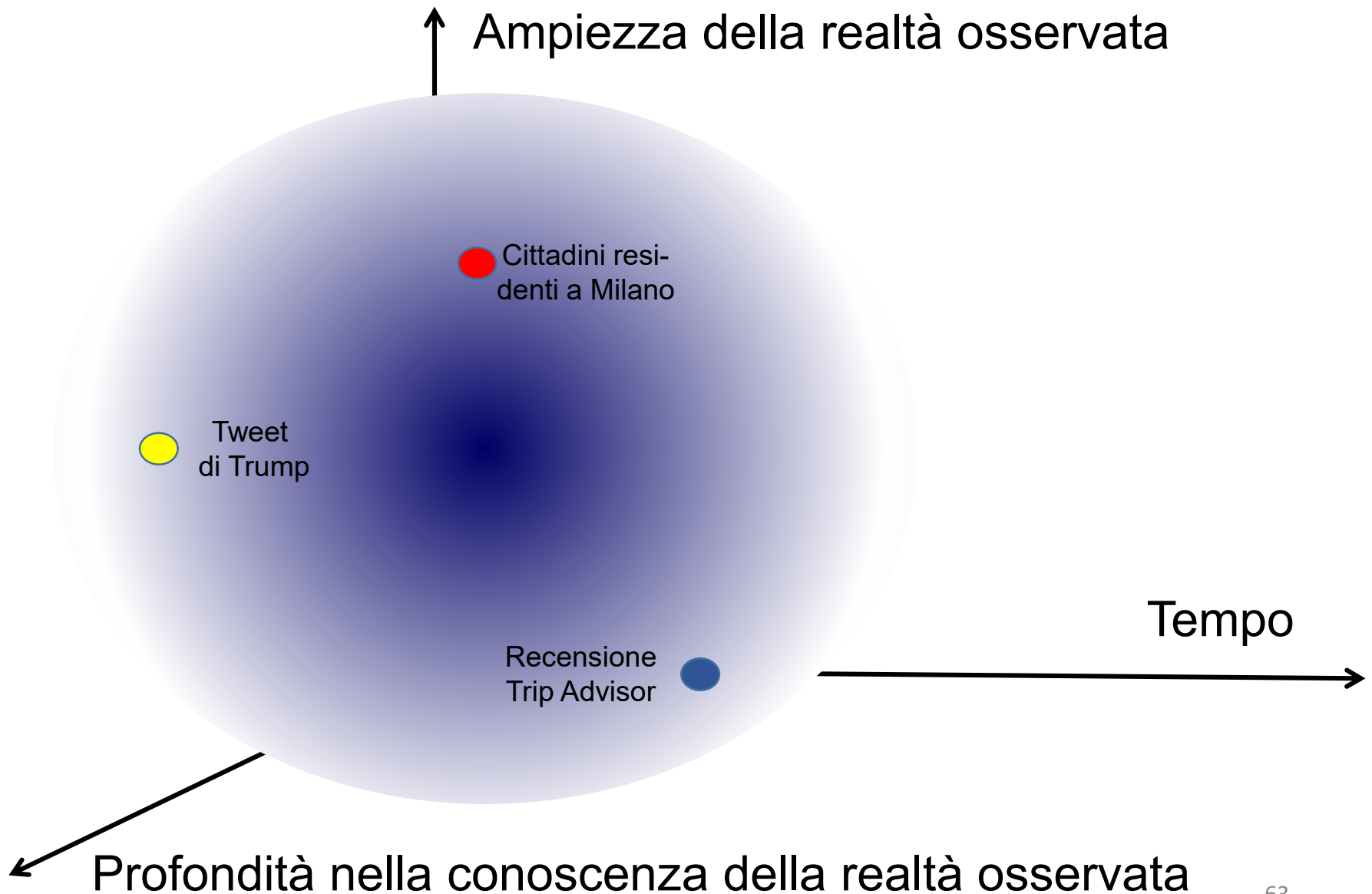
Dear Laure, I try to describe the wonderful harbour of Portofino as I have seen this morning a boat is going in, other boats are along the wharf. Small pretty buildings and villas are looking on to the harbour.



Principali caratteristiche dei Big Data

- **Volume**, la qualità di dati che possono essere raccolti, memorizzati, elaborati
- **Velocità**, la velocità con cui i dati cambiano
- **Varietà**, la diversificazione del formato dei dati, da strutturati a non strutturati (testi) a immagini
- **Veridicità**, quanto i dati aderiscono accuratamente al fenomeno che descrivono

La grande sfera opaca



Alberto Mantovani

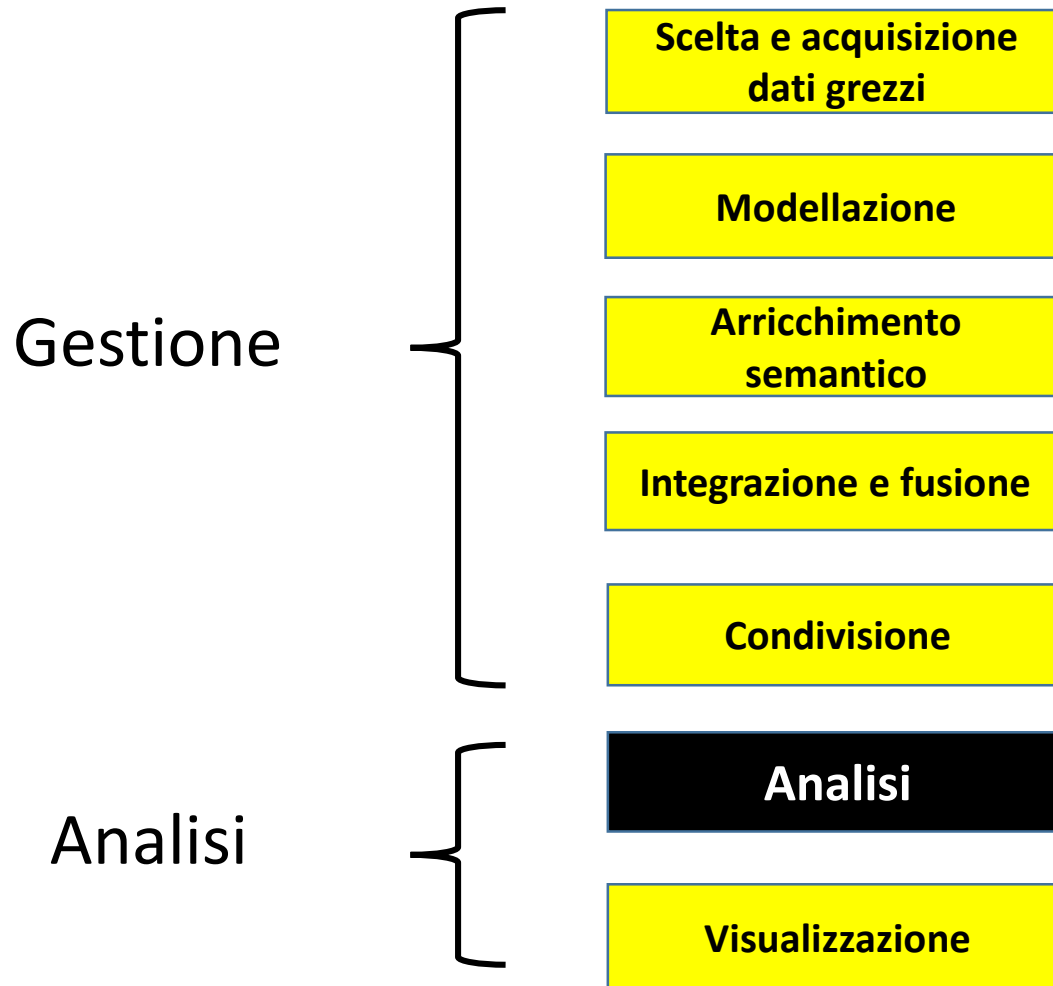
Non avere paura di sognare: Decalogo per aspiranti scienziati



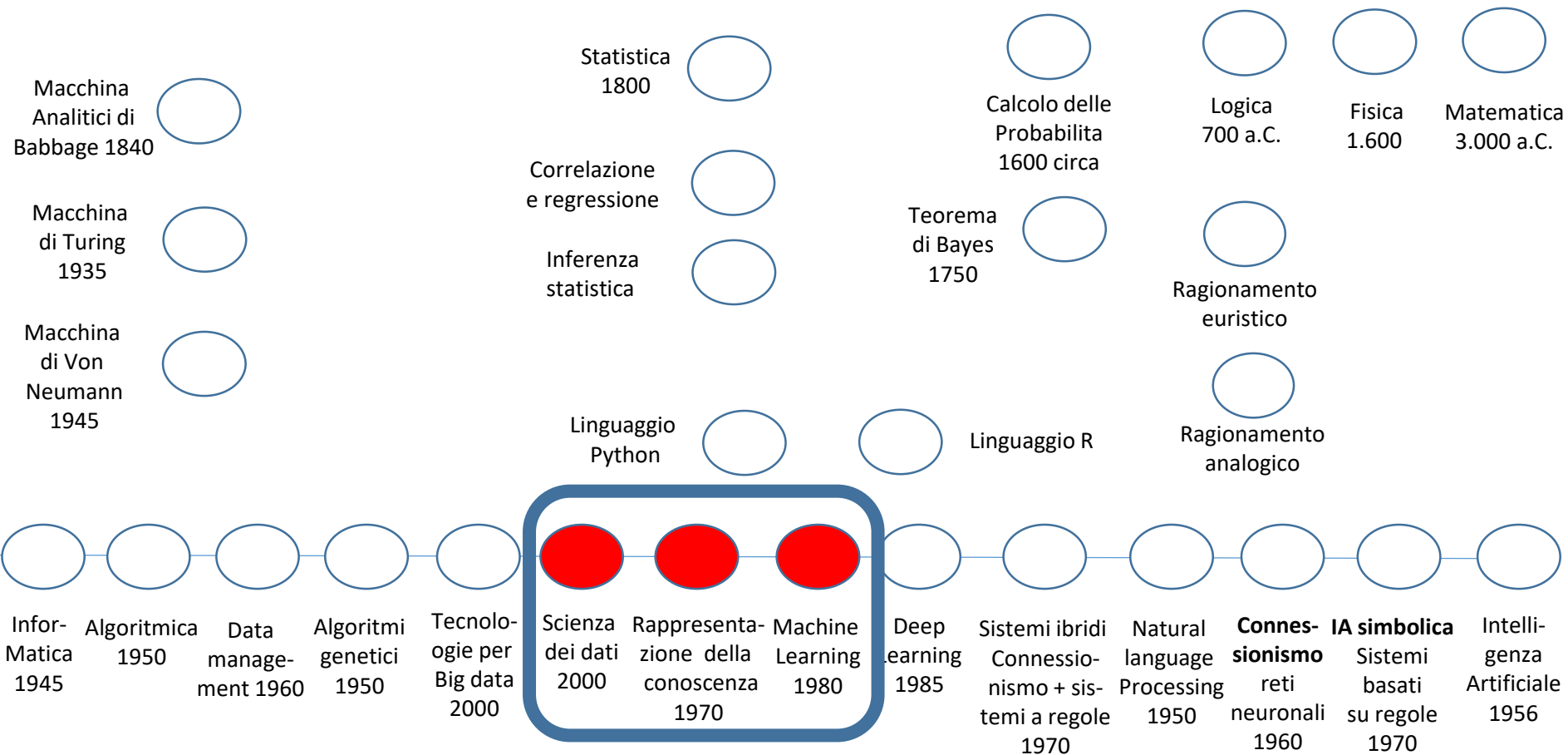
Introduzione

1. Segui le tue passioni
2. Vivi in una dimensione internazionale e contribuisci a costruire ponti di pace
3. Sii umile
4. Raccogli le sfide
5. Impara dai pazienti
6. Collabora e sii pronto ad ascoltare gli altri
7. Impara dai tecnici: la chiave a stella
8. Accetta il giudizio degli altri e fatti guidare dal tuo spirito critico
9. Rispetta i dati
10. Condividi e cambia il mondo

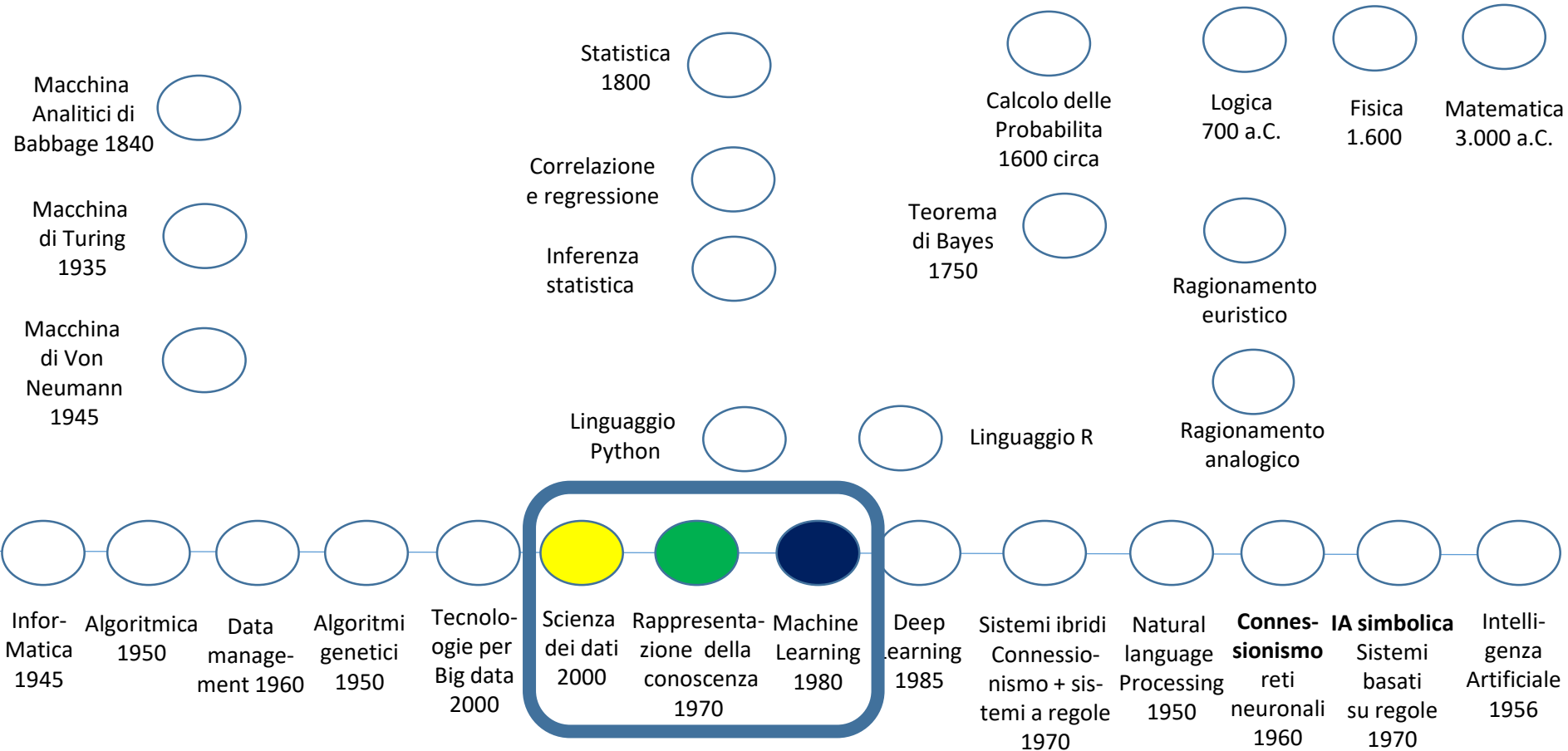
Bisogna aver cura dei dati il ciclo di vita del dato



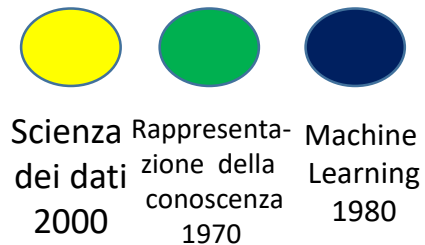
Genesi e sviluppi della Scienza dei Dati



Genesi e sviluppi della Scienza dei Dati



La Scienza dei dati in sintesi



Ciclo di vita del dato

**Scelta e acquisizione
dati grezzi**

Modellazione

**Arricchimento
semantico**

Integrazione e fusione

Condivisione

Analisi

Visualizzazione

I tipi di analisi

Analisi



Analisi
descrittive

Cosa è la realtà?



Analisi
interpretative

Perchè è accaduto?



Analisi
predittive

Cosa sarà la realtà?



Analisi
prescrittive

Come intervenire
sulla realtà ?

Esempio di Analisi predittiva

Prevedere l'inquinamento



Prevedere l'inquinamento



Breezometer



Quale sarà il livello
di inquinamento
domani a Milano
a Via Tortona alle 15?



Scelta e acquisizione dati grezzi

Modellazione

Arricchimento

Integrazione e fusione

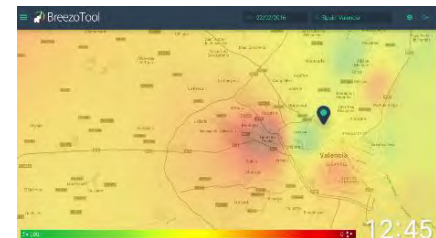
Condivisione

Analisi

Visualizzazione



Problema:
Prevedere
l'inquinamento



Esempio di analisi descrittiva

Fare statistiche sui collaboratori di una azienda

Miroslav	Konecny	1978	7	10	Street	St. John	49	Prague	412776	null
Martin	Necasky	1975	7	8	Sq.	Vienna	null	Bratislava	101278	Slovensko
Miroslav	Konecny	null	null	null	Str.	Saint Jon	49	Prague	412776	null
Crlo	Btini	1949	June	7	Street	Dessiè	15	Roma	00198	Italia
Miroslav	Knecy	1978	7	10	Sq.	Budapest	23	Wien	null	Austria
Anisa	Rula	1982	September	7	Via	Sesto	null	Milano	20...	Ital
Anita	Rula	1982	9	7	Via	Seto	23	Milan	null	Italy
Anna	Rla	null	null	null	Via	Sarca	336	Milano	null	Italy
Carlo	Batini	1949	6	7	V.	Beato Angelico	23	Milan	20133	Italy
Carla	Botni	1949	June	7	Av.	Charles	null	Prague	412733	null
Marta	Necasky	1976	11	8	null	null	null	Bratislava	112234	Slov.

Maria	Roma	1978	7	10	St. John	40	Prog	412	null
Mark	Nice	1978	7	8	St.	Vienna	null	Stell	101
Maria	Roma	null	null	null	St.	St. John	40	Prog	412
Clara	Stell	1984	June	7	St. John	15	Prog	001	Italy
Maria	Nice	1978	7	10	St.	Budapest	23	Wien	null
Anna	Roma	1982	Sept	7	Via	Stell	null	Milan	20
Anna	Roma	1982	8	7	Via	Stell	23	Milan	null
Anna	St.	null	null	null	Via	Stell	230	Milan	null
Carlo	Stell	1984	8	6	V.	Stell	23	Milan	201
Carlo	Stell	1984	8	June	Av.	Charles	null	Prog	412
Mark	Nice	1978	11	8	null	null	null	Stell	112



Scelta e acquisizione dati grezzi

Modellazione

Arricchimento

Integrazione e fusione

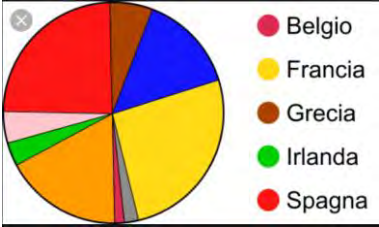
Condivisione

Analisi

Visualizzazione



Problema:
Fare statistiche
sui collaboratori



Vediamo ora il ciclo di vita del dato digitale
fase per fase

Scelta e acquisizione dati grezzi

Modellazione

Arricchimento semantico

Integrazione e fusione

Condivisione

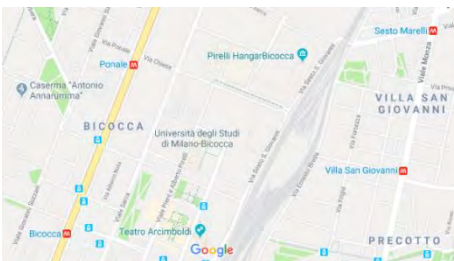
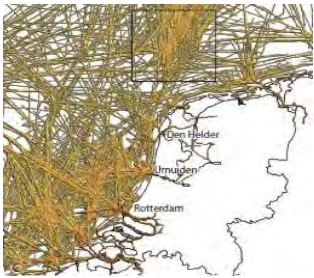
Analisi

Visualizzazione

I dati vanno trovati,
scelti e acquisiti

Prevedere l'inquinamento

Scelta e acquisizione dati



Dati satellitari

**Fonti di dati
ambientali**

Flussi di trasporto

Sensori nei Comuni

Google Maps

Scelta e acquisizione dati grezzi

Modellazione

Arricchimento semantico

Integrazione e fusione

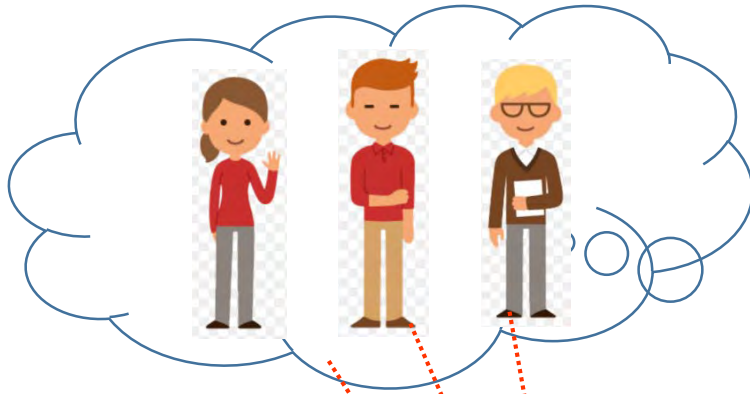
Condivisione

Analisi

Visualizzazione

I dati vanno descritti
con modelli

Tre milanesi



Vogliamo rappresentare tre studenti della Università di Milano-Bicocca con Matricola, Cognome, Comune di Nascita, Data di Nascita, Indirizzo di residenza e CAP

Cittadini Residenti a Milano

Matricola	Cognome	Comune di Nascita	Data di nascita	Indirizzo di residenza	CAP
32127	Rossi	Pescara	7/6/1949	Via Rossini 18	20127
17843	Batini	Roma	9/7/1999	Via Rossini 18	20127
23627	Rossi	Milano	20/12/1996	Via Verdi 31	20109

La tabella, le righe e gli attributi

Studenti della Università di Milano Bicocca

Attributo



Matricola	Cognome	Comune di Nascita	Data di nascita	Indirizzo di residenza	CAP
32127	Rossi	Pescara	7/6/1949	Via Rossini 18	20127
17843	Batini	Roma	9/7/1999	Via Rossini 18	20127
23627	Rossi	Milano	20/12/1996	Via Verdi 31	20109

Riga



Schema e istanza

Schema

Studenti della Università di Milano-Bicocca

Matricola	Cognome	Comune di Nascita	Data di nascita	Indirizzo di residenza	CAP
32127	Rossi	Pescara	7/6/1949	Via Rossini 18	20127
17843	Batini	Roma	9/7/1999	Via Rossini 18	20127
23627	Rossi	Milano	20/12/1996	Via Verdi 31	20109

Istanza

Il linguaggio SQL

Studente

Matricola	Nome	Cognome
13242	Carlo	Batini
24195	Ling	Xu
32845	John	Smith

Esame

Matricola	CodiceCorso	Voto
13242	27	25
24195	77	28
32845	27	30

Corso

Codice	NomeCorso	Anno
27	Analisi	1
49	Algoritmi	1
77	Logica	2

**Trova i codici degli esami superati da Batini
e il voto ottenuto, ma solo quelli con voto > 28**

↓
Linguaggio SQL
↓

```
SELECT Esame.CodiceCorso, Voto
FROM Studente, Esame
WHERE Studente.Matricola = Esame.CodiceCorso AND Voto > 28
```

Scelta e acquisizione dati grezzi

Modellazione

Arricchimento semantico

Integrazione e fusione

Condivisione

Analisi

Visualizzazione

I dati vanno
arricchiti

Scelta e acquisizione dati grezzi

Modellazione

Arricchimento semantico

Integrazione e fusione

Condivisione

Analisi

Visualizzazione

Cura della qualità

Modelli più ricchi

Scelta e acquisizione dati grezzi

Modellazione

Arricchimento semantico

Integrazione e fusione

Condivisione

Analisi

Visualizzazione

Cura della qualità

Un esempio di dati di scarsa qualità

27ª Giornata — 03/03/2013

Catania 2 • 4 **Inter**

MARCHESE (19° 1T) BERGESSIO (7° 1T) PALACIO (48° 2T) PALACIO (47° 2T) PALACIO (25° 2T) ALVAREZ (7° 2T)

CATANIA 90 1T 5 10 15 20 25 30 35 40 45 2T 5 10 15 20 25 30 35 40 45

INTER

TABELLINO

Catania	Inter
21 ANDUJAR	1 HANDANOVIC
5 ROLIN	40 JUAN JESUS
12 MARCHESE	31 PEREIRA
3 SPOLLI	26 CHIVU
22 ALVAREZ	11 ALVAREZ
27 BIAGIANTI	14 GUARIN
13 IZCO	4 ZANETTI
10 LODI	22 KUZMANOVIC
19 CASTRO	21 GARGANO
17 GOMEZ	7 SCHELOTTO

COMMENTO IN DIRETTA di Pasquale Notargiacomo

Primo Tempo **Secondo Tempo**

49° **FINISCE LA PARTITA! Vince l'Inter 3-2**

48° **IL GOL. Cambiasso in area, vicino alla linea di fondo, serve un assist perfetto per il rasoterra in area di Palacio che completa la rimonta**

47° **INTER IN VANTAGGIO!!!! 3-2**

46° Destro di Almiron da fuori area: potente ma impreciso

45° Saranno quattro i minuti di recupero

44° **L'INTER SFIORA IL TERZO GOL! Diagonale riavvicinato di Cambiasso in area, fuori di**

DIRETTE **RISULTATI**

Calcio - Serie A

03.03 Torino - Palermo	0 - 0	12:30
03.03 Bologna - Cagliari	2 - 0	15:00
03.03 Catania - Inter	2 - 2	15:00
03.03 Fiorentina - Chievo	2 - 1	15:00
03.03 Pescara - Udinese	0 - 1	15:00
03.03 Sampdoria - Parma	1 - 0	15:00
03.03 Siena - Atalanta	0 - 2	15:00
03.03 Roma - Genoa	0 - 0	20:45

Legenda: ⌚ In programma 👁 Diretta 🚫 Finita

Serie A **Marcatori**

Quale di queste due immagini
e' di migliore qualità?



Errori in una Anagrafe

Nome	Cognome	Indirizzo	Comune
Crlo	Batini	Piazzale Lreto 3	Mlano
Carla	Neri	Via Monte Velino 2	Rocca di Mezzo

Quale di queste foto di Portofino
è di migliore qualità?
Dipende...se c'era una nebbiolina....

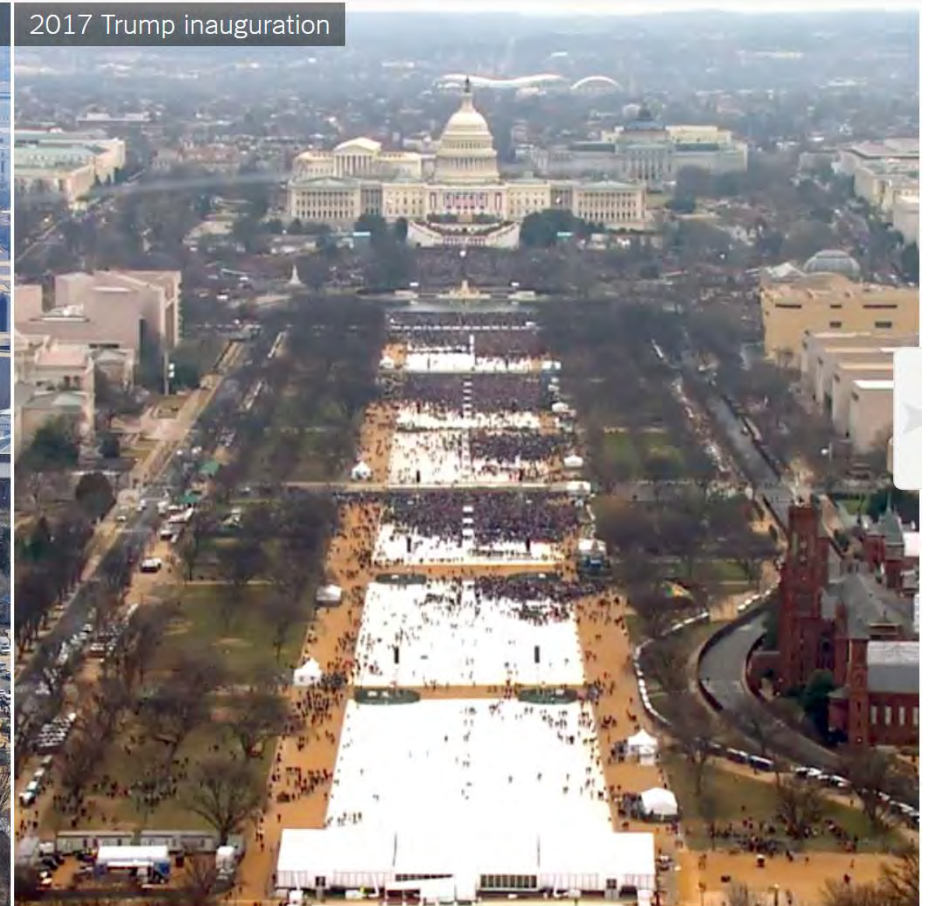
Più precisa



Più fedele



Kelianne Conway, le cerimonie di insediamento di Obama e di Trump e i «fatti alternativi»



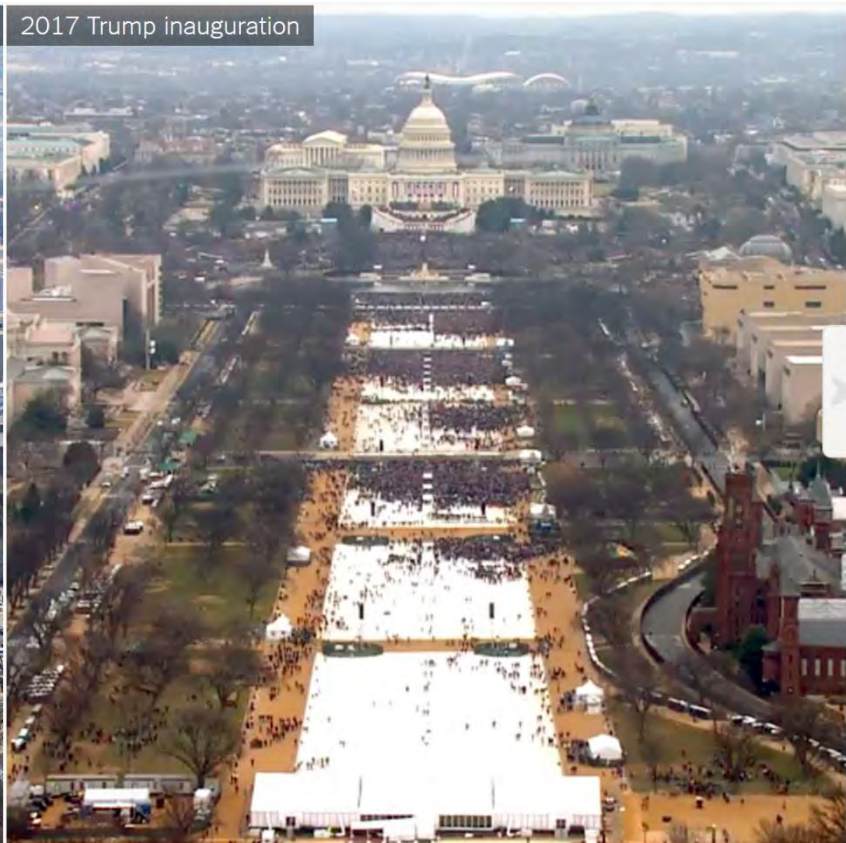
Come possiamo capire chi ha ragione? I fatti sono testardi



Ora della foto: 11.30

Ora della cerimonia: 12

**Numero biglietti metropolitana
nell'ora precedente: 80.000**

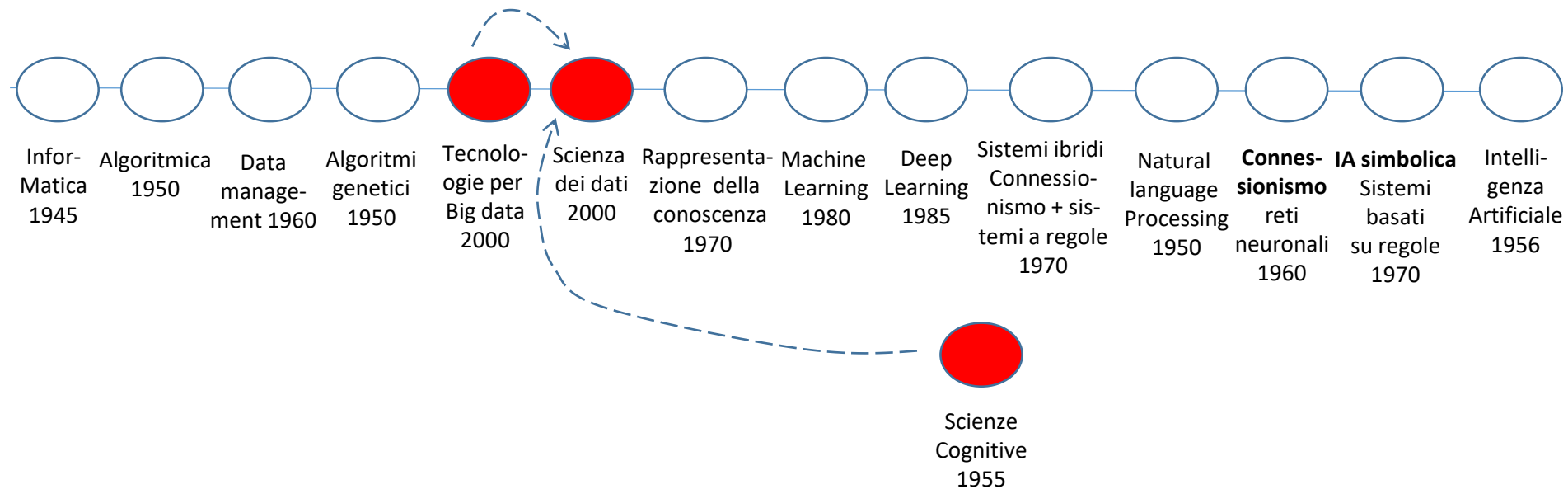


Ora della foto: 11.30

Ora della cerimonia: 12

**Numero biglietti metropolitana
nell'ora precedente: 20.000**

Le scienze cognitive e le reti sociali



Dalle scienze cognitive non arrivano buone notizie

**Non è tanto
rilevante ciò che
la gente pensa,
ma come pensa.
Riconoscere la
cattiva
informazione
richiede processi
cognitive
complessi**

**Un semplice
mito è più
attraattivo
cognitiva-
mente di
una
complicata
correzione**

**Per coloro che
sono
fortemente
convinti delle
proprie idee,
gli argomenti
fortemente
contrari
possono
rafforzare le
loro convinzioni**

Scelta e acquisizione dati grezzi

Modellazione

Arricchimento semantico

Integrazione e fusione

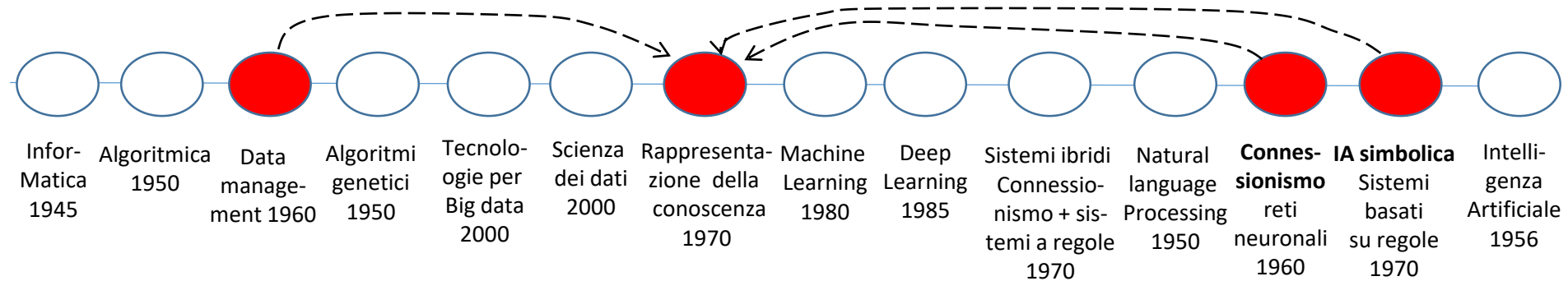
Condivisione

Analisi

Visualizzazione

Modelli più ricchi

La rappresentazione della conoscenza



1. Contributi da Informatica

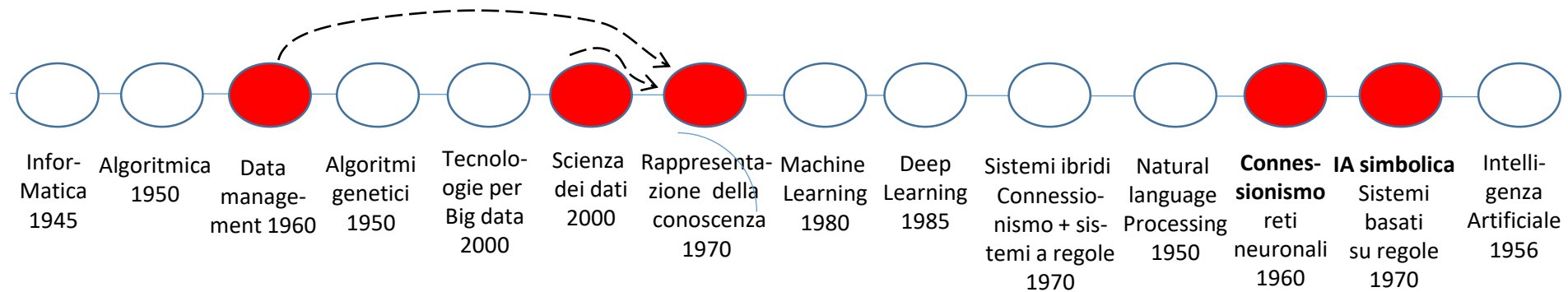


Tabella di partenza: cosa significano le varie colonne?

Collaboratori

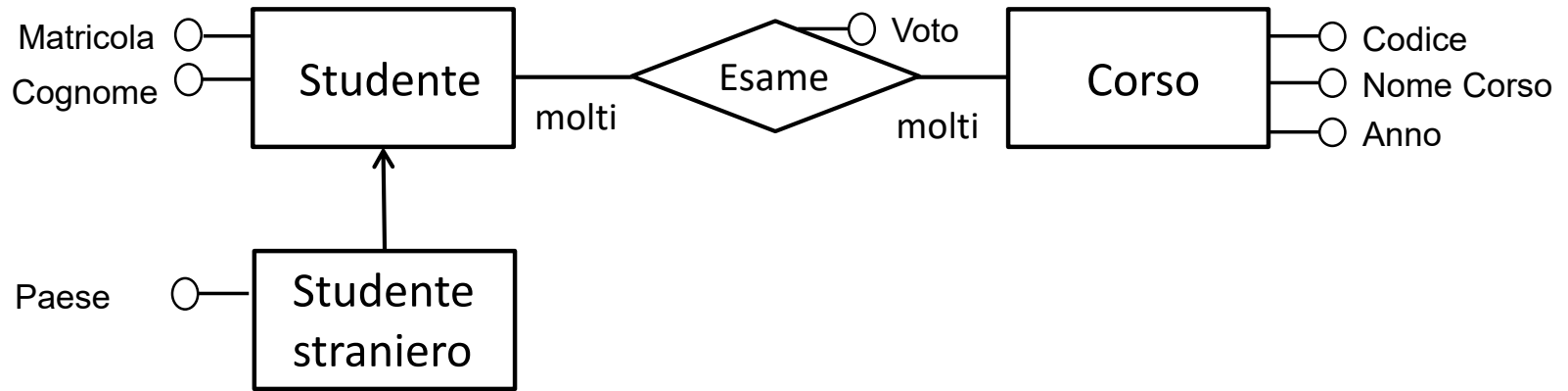
Miroslav	Konecny	1978	7	10	Street	St. John	49	Prague	412776	null
Martin	Necasky	1975	7	8	Sq.	Vienna	null	Bratislava	101278	Slovensko
Miroslav	Konecny	null	null	null	Str.	Saint Jon	49	Prague	412776	null
Crlo	Btini	1949	June	7	Street	Dessiè	15	Roma	00198	Italia
Miroslav	Knecy	1978	7	10	Sq.	Budapest	23	Wien	null	Austria
Anisa	Rula	1982	September	7	Via	Sesto	null	Milano	20...	Ital
Anita	Rula	1982	9	7	Via	Seto	23	Milan	null	Italy
Anna	Rla	null	null	null	Via	Sarca	336	Milano	null	Italy
Carlo	Batini	1949	6	7	V.	Beato Angelico	23	Milan	20133	Italy
Carla	Botni	1949	June	7	Av.	Charles	null	Prague	412733	null
Marta	Necasky	1976	11	8	null	null	null	Bratislava	112234	Slov.

Nomi per le colonne

Collaboratori

Tuple #	Nome	Cognome	Anno di nascita	Mese di Nascita	Giorno nascita	Toponimo	Nome del Toponimo	Numero Civico	Città	CAP	Paese
1	Miroslav	Konecny	1978	7	10	Street	St. John	49	Prague	412776	null
2	Martin	Necasky	1975	7	8	Sq.	Vienna	null	Bratislava	101278	Slovensko
3	Miroslav	Konecny	null	null	null	Str.	Saint Jon	49	Prague	412776	null
4	Crlo	Btini	1949	June	7	Street	Dessiè	15	Roma	null	Italia
5	Miroslav	Knecy	1978	7	10	Sq.	Budapest	23	Wien	null	Austria
6	Anisa	Rula	1982	September	7	Via	Sesto	null	Milano	20...	Ital
7	Anita	Rula	1982	9	7	Via	Seto	23	Milan	null	Italy
8	Anna	Rla	null	null	null	Via	Sarca	336	Milano	null	Italy
9	Carlo	Batini	1949	6	7	V.	Beato Angelico	23	Milan	20133	Italy
10	Carla	Botni	1949	June	7	Av.	Charles	null	Prague	412733	null
11	Marta	Necasky	1976	11	8	null	null	null	Bratislava	112234	Slov.

Dal modello relazionale al modello Entità Relazione



Studente

Matricola	Cognome	Paese
13242	Batini	-
24195	Xu	Cina
32845	Smith	USA

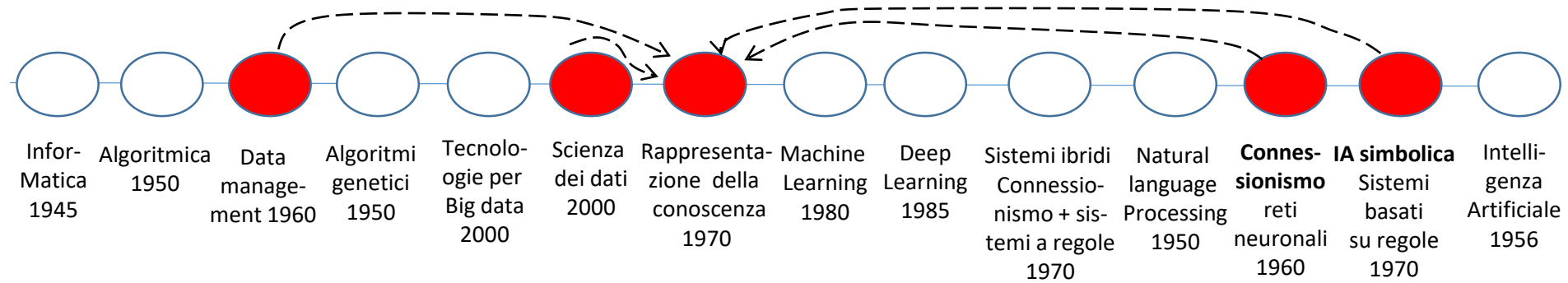
Esame

Matricola	CodiceCorso	Voto
13242	27	25
24195	77	28
32845	27	30

Corso

Codice	NomeCorso	Anno
27	Analisi	1
49	Algoritmi	1
77	Logica	2

2. Contributi “misti”



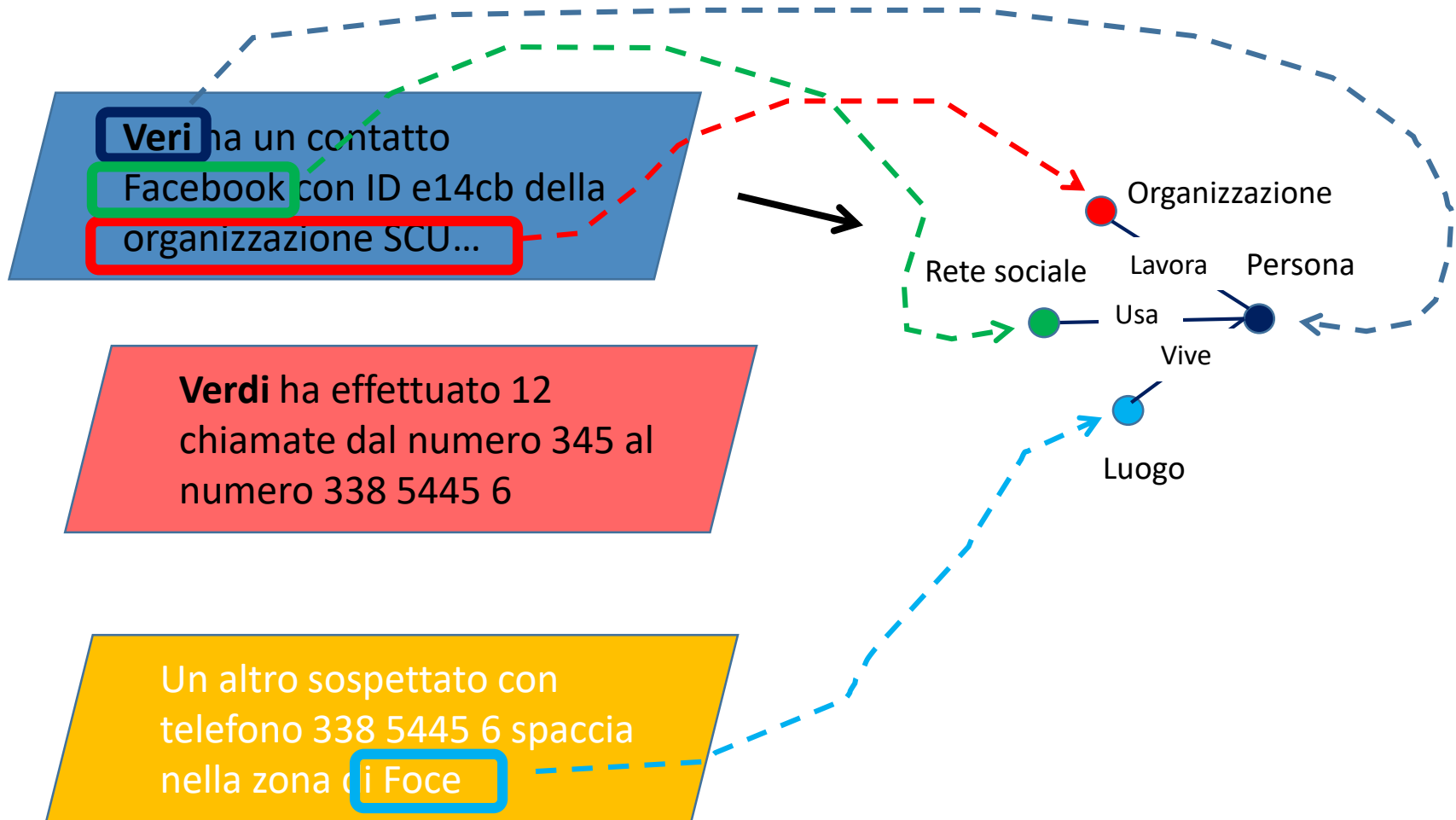
Esempio di Indagine dallo sfogliare documenti...

Veri ha un contatto
Facebook con ID e14cb della
organizzazione SCU...

Verdi ha effettuato 12
chiamate dal numero 345 al
numero 338 5445 6

Un altro sospettato con
telefono 338 5445 6 spaccia
nella zona di Foce

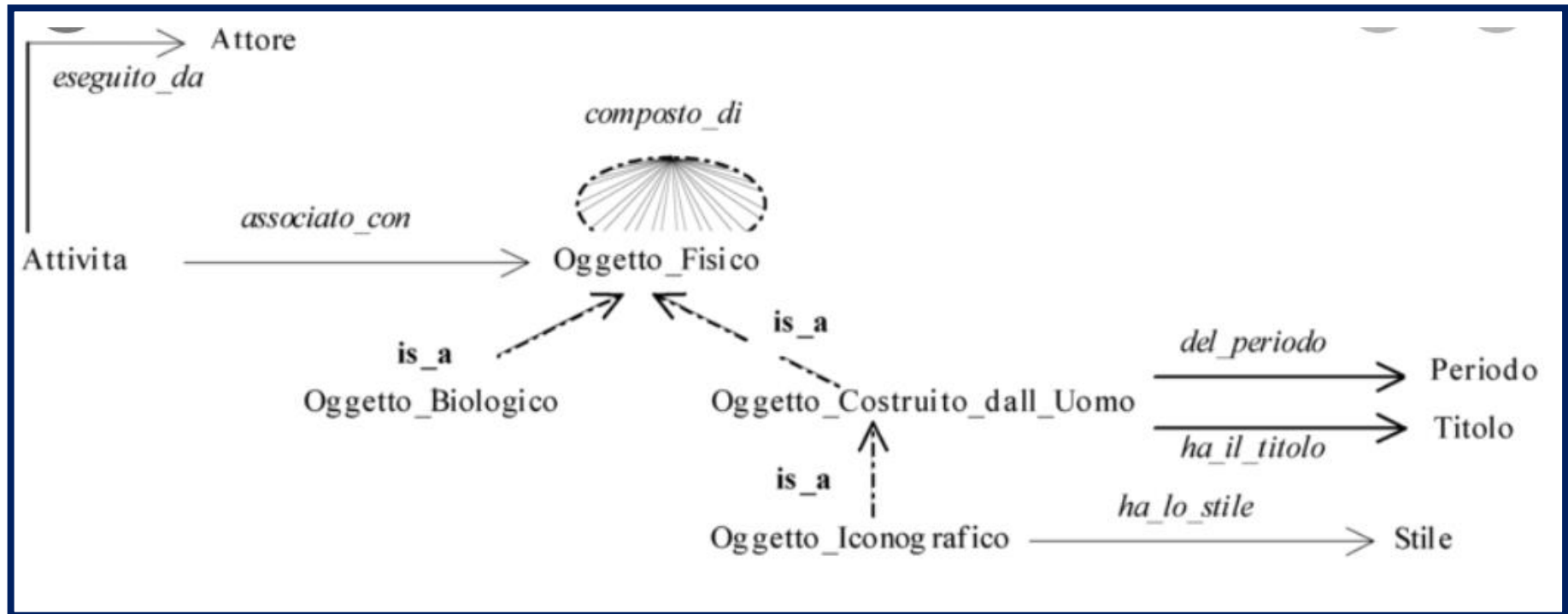
..... alla integrazione dei dati in un grafo



Novità - 1

- Rappresentazione a grafo
- La semantica dei concetti è formale (es. **Grafi semantici, ontologie**) e quindi **elaborabile da una macchina**)

Esempio di Ontologia in Biologia



Novità - 2

- Disponibilità di risorse semantiche sul Web (es. Wikipedia)
- Molte funzioni di disambiguazione, matching logico, matching spaziale, matching temporale, esplorazione, interrogazione, georeferenziazione, individuazione di anomalie, ecc.

Esempio di uso di risorsa semantica per disambiguazione

- Verdi è stato **individuato** a Palermo,
- Presso un ristorante che **vendeva** Asado
- Vicino alla **Vucciria**

Palermo

Da Wikipedia, l'enciclopedia libera.

Disambiguazione – Se stai cercando altri significati, vedi **Palermo (disambigua)**.

Questa voce o sezione ha problemi di struttura e di organizzazione delle informazioni.

Motivo: Voce dalle dimensioni abnormi, oltre il doppio delle dimensioni massime consigliate. Bisognerebbe controllare gli ancillari, probabili alcune ridondanze ed eccessivo ipertesto.

Risistema la struttura espositiva, logica e/o bibliografica dei contenuti. Nella **discussione** puoi collaborare con altri utenti.

Palermo (AFI: [paˈlɛrmo]^[5] ascolta^[?·info]; *Palermu* [paˈlɛmmu] in siciliano, *Paliemmu* in dialetto palermitano^[6]) è un comune italiano di 663 770 abitanti^[3], capoluogo dell'omonima città metropolitana e della Regione siciliana. È il quinto comune italiano per popolazione nonché il principale centro urbano dell'isola di Sicilia e dell'Italia insulare, secondo nel Mezzogiorno d'Italia e trentesimo a livello europeo.

La città vanta una storia plurimillenaria e ha avuto un ruolo importante per le vicende del Mediterraneo e dell'Europa. Fondata dai Fenici tra il VII e il VI secolo a.C.^[7], viene conquistata nel 254 a.C. dai Romani ed è divenuta il principale centro dell'isola (Provincia Sicilia)^[8]. Conquistata dai Vandali nel 429, passa nel 536 sotto Bisanzio e poi, brevemente, agli Arabi, che la conquistano nel 647. Successivamente, con l'avvento dei Normanni e il ritorno alla cristianità, fu la città di riferimento per i numerosi sovrani di Sicilia, proprio a questa circostanza si devono i titoli attribuiti alla città: *Prima Sedes*, *Corona Regis et Regni Caput*. Da allora è rimasta, con alterne vicende, la capitale del Regno di Sicilia fino al 1816. Particolarmente importanti furono i Vespri siciliani, ribellione scoppiata a Palermo nel 1292.

Palermo (Buenos Aires)

Da Wikipedia, l'enciclopedia libera.

Palermo è un **barrio** di Buenos Aires, capitale dell'Argentina. Sorge nella zona nord-est della città e confina a nord con Belgrano, a est con il Río de la Plata, a sud con Recoleta e Almagro e ad ovest con Villa Crespo e Colegiales. Con una superficie di 15.4 km² è il quartiere più esteso di tutta la città. È composto principalmente da edifici residenziali, che sono stati sempre più negli ultimi dieci anni. Ospita il grande Parque Tres de Febrero, con il Planetario Galileo Galilei.

Indice [nascondi]

- 1 Etimologia
- 2 Suddivisione interna
- 3 Trasporto pubblico
- 4 Altri progetti

Etimologia [modifica] [modifica wikitesto]

Il nome del quartiere deriva da un convento non più esistente dedicato a San **Benedetto** da San Fratello, conosciuto anche come San Benedetto da Palermo.

Scelta e acquisizione dati grezzi

Modellazione

Arricchimento semantico

Integrazione e fusione

Condivisione

Analisi

Visualizzazione

I dati vanno
ri-messi insieme

Scelta e acquisizione dati grezzi

Modellazione

Arricchimento semantico

Integrazione e fusione

Condivisione

Analisi

Visualizzazione

Integrazione

Fusione

Scelta e acquisizione dati grezzi

Modellazione

Arricchimento semantico

Integrazione e fusione

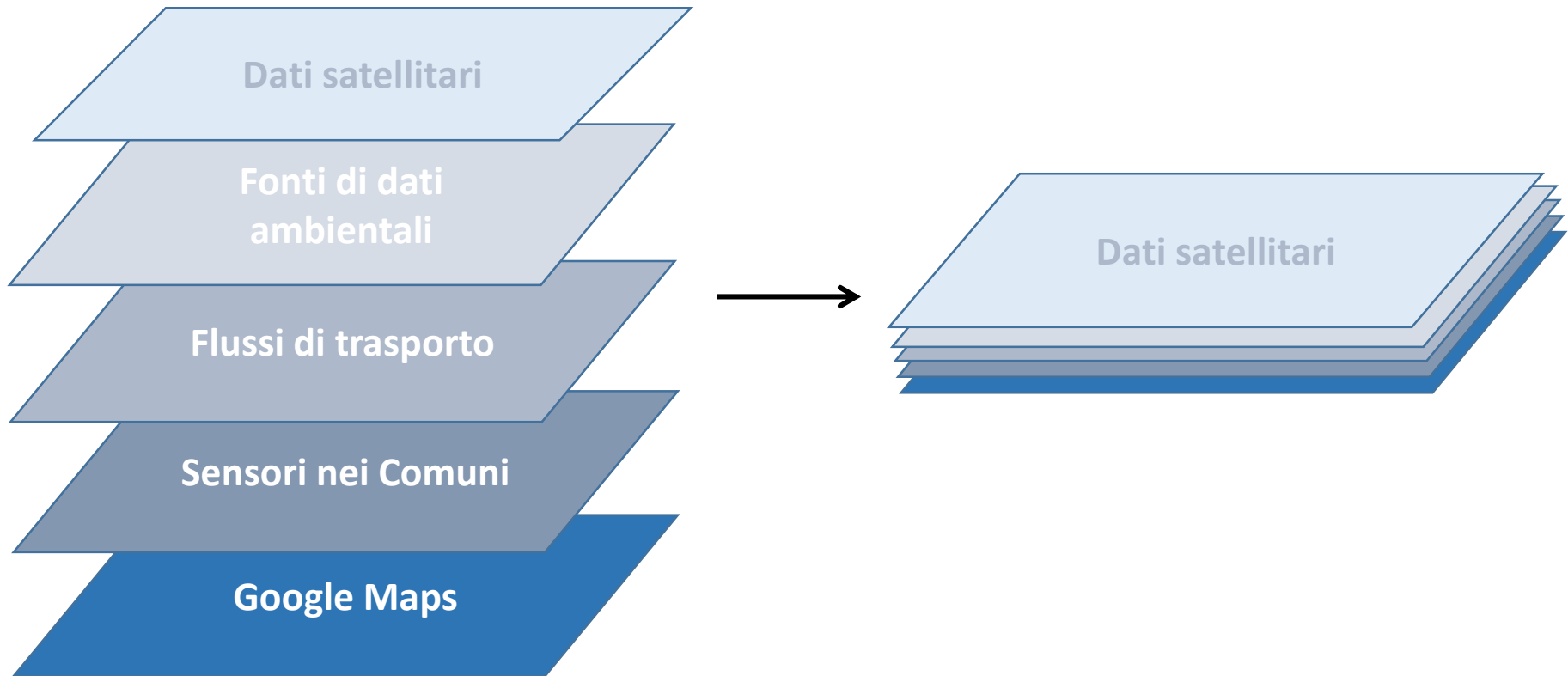
Condivisione

Analisi

Visualizzazione

Integrazione

Integrazione nel problema sull'inquinamento



Esempio 1

stesso identificatore → CF

Condizione → stesso valore per CF

CF	Nome	Cognome	Città N.
CRLBTN	Carlo	Batini	Roma
GNTRSS	Gino	Rossi	Venezia
LDTBSS	Aldo	Busi	Lodi

CF	Nome	Cognome	Città N.
CRLBTN	Carlo	Batini	Roma
GTTRND	Gino	Rossi	Milano
LDTBSS	Aldo	Bui	Lodi

Esempio 2

Diverso identificatore → CF vs Matricola

Condizione → Nome e Cognome identico

CF	Nome	Cognome	Città N.
CRLBTN	Carlo	Batini	Roma
GNTRSS	Gino	Rossi	Venezia
LDTBSS	Aldo	Busi	Lodi

Matricola	Nome	Cognome	Città N.
03475	Carlo	Batini	Roma
43287	Gino	Rossi	Venezia
78125	Aldo	Bui	Lodi

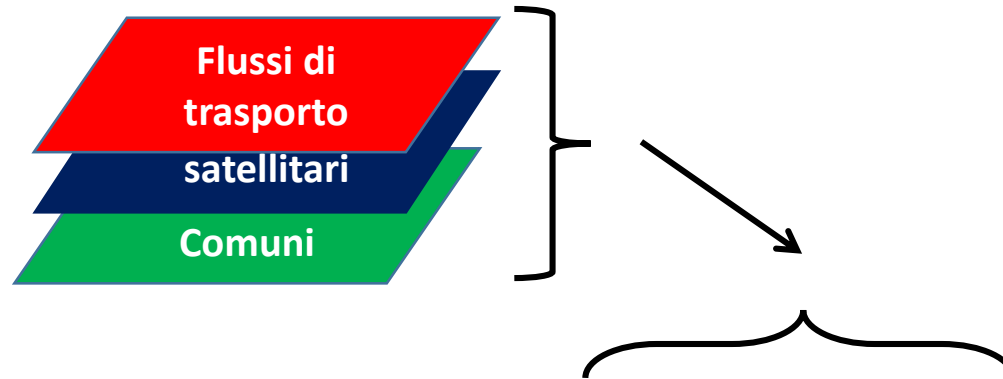
Esempio 3

Condizione → Nome e cognome differiscono di al massimo un carattere

CF	Nome	Cognome	Città N.
CRLBTN	Carlo	Batini	Roma
GNTRSS	Gino	Rossi	Venezia
LDTBSS	Aldo	Busi	Lodi

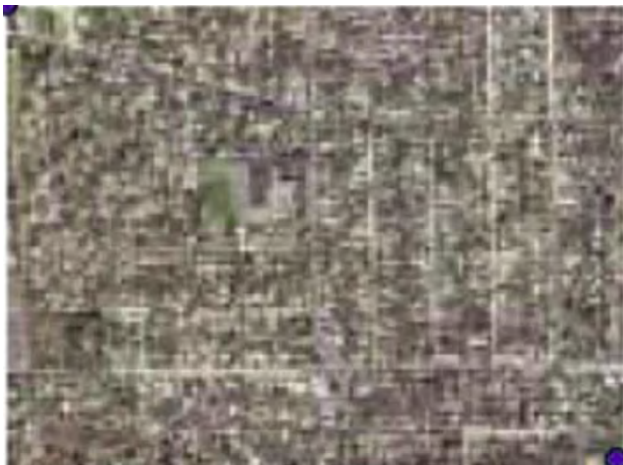
Matricola	Nome	Cognome	Città N.
03475	Carlo	Batini	Roma
43287	Gino	Rossi	Venezia
78125	Aldo	Bui	Lodi

Inquinamento



latitudine	longitudine	Anno	Mese	Giorno	Ora	Minuto	Traffico	Riscaldamento	Qualità dell'aria	Livello di inquinamento
308.40	215.20	2017	6	12	8	10	0.7	0.7	47	3
308.40	215.20	2017	7	12	8	20	0.7	0.8	49	2
308.40	215.20	2017	6	12	8	30	0.8	0.8	51	5
308.40	215.20	2017	6	12	8	40	0.9	0.7	54	6
..	..	..	..	..	..	.			..	

L'integrazione nelle mappe



Scelta e acquisizione dati grezzi

Modellazione

Arricchimento semantico

Integrazione e fusione

Condivisione

Analisi

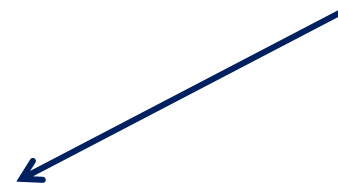
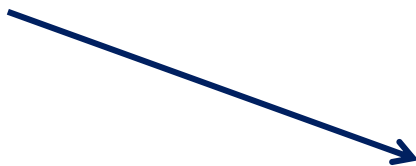
Visualizzazione

Fusione

Fusione

CF	Nome	Cognome	Città N.	Anno N.
CRLBTN	Carlo	Batini	Roma	1976
GNTRSS	Gino	Rossi	Venezia	1999
LDTBSS	Aldo	Busi	Lodi	2001

CF	Nome	Cognome	Città N.	Anno N.
CRLBTN	Carlo	Batini	Roma	1978
GNTRSS	Gino	Rossi	Milano	1999
LDTBSS	Aldo	Busi	Lodi	2001



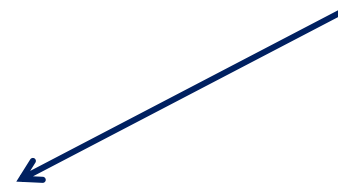
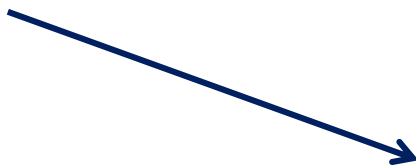
CF	Nome	Cognome	Città N.	Anno N.
CRLBTN	Carlo	Batini	Roma	??
GNTRSS	Gino	Rossi	??	1999
LDTBSS	Aldo	Busi	Lodi	2001

Per l' Anno di Nascita – Strategia 1

Inserisci la media

CF	Nome	Cognome	Città N.	Anno N.
CRLBTN	Carlo	Batini	Roma	1976
GNTRSS	Gino	Rossi	Venezia	1999
LDTBSS	Aldo	Busi	Lodi	2001

CF	Nome	Cognome	Città N.	Anno N.
CRLBTN	Carlo	Batini	Roma	1978
GNTRSS	Gino	Rossi	Milano	1999
LDTBSS	Aldo	Busi	Lodi	2001

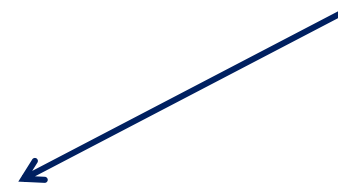


CF	Nome	Cognome	Città N.	Anno N.
CRLBTN	Carlo	Batini	Roma	1977
GNTRSS	Gino	Rossi	??	1999
LDTBSS	Aldo	Busi	Lodi	2001

Per l' Anno di Nascita – Strategia 2 inseriscile tutte e due

CF	Nome	Cognome	Città N.	Anno N.
CRLBTN	Carlo	Batini	Roma	1976
GNTRSS	Gino	Rossi	Venezia	1999
LDTBSS	Aldo	Busi	Lodi	2001

CF	Nome	Cognome	Città N.	Anno N.
CRLBTN	Carlo	Batini	Roma	1978
GNTRSS	Gino	Rossi	Milano	1999
LDTBSS	Aldo	Busi	Lodi	2001



CF	Nome	Cognome	Città N.	Anno N.
CRLBTN	Carlo	Batini	Roma	1976-1978
GNTRSS	Gino	Rossi	??	1999
LDTBSS	Aldo	Busi	Lodi	2001

E per la città di Nascita?

CF	Nome	Cognome	Città N.	Anno N.
CRLBTN	Carlo	Batini	Roma	1976
GNTRSS	Gino	Rossi	Venezia	1999
LDTBSS	Aldo	Busi	Lodi	2001

CF	Nome	Cognome	Città N.	Anno N.
CRLBTN	Carlo	Batini	Roma	1978
GNTRSS	Gino	Rossi	Milano	1999
LDTBSS	Aldo	Busi	Lodi	2001

Per esempio → Scegliere quella inserita più recentemente

Individuiamo i duplicati nella tabella dei professionisti
e fondiamo i duplicati in una unica riga

Tuple #	Nome	Cognome	Anno di nascita	Mese di Nascita	Giorno nascita	Toponimo	Nome del Toponimo	Numero Civico	Città	CAP	Paese
1	Miroslav	Konecny	1978	7	10	Street	St. John	49	Prague	412776	null
2	Martin	Necasky	1975	7	8	Sq.	Vienna	null	Bratislava	101278	Slovensko
3	Miroslav	Konecny	null	null	null	Str.	Saint Jon	49	Prague	412776	null
4	Crlo	Btini	1949	June	7	Street	Dessiè	15	Roma	null	Italia
5	Miroslav	Knecy	1978	7	10	Sq.	Budapest	23	Wien	null	Austria
6	Anisa	Rula	1982	September	7	Via	Sesto	null	Milano	20...	Ital
7	Anita	Rula	1982	9	7	Via	Seto	23	Milan	null	Italy
8	Anna	Rla	null	null	null	Via	Sarca	336	Milano	null	Italy
9	Carlo	Batini	1949	6	7	V.	Beato Angelico	23	Milan	20133	Italy
10	Carla	Botni	1949	June	7	Av.	Charles	null	Prague	412733	null
11	Marta	Necasky	1976	11	8	null	null	null	Bratislava	112234	Slov.

Tabella finale per professionisti con eliminazione dei duplicati

Tuple #	Candidate tuples	F.Name	Last Name	Y.Of Birth	M. Birth	D. Birth	Toponym	Name Toponym	Number T.	City	Zip Code	Country
1	M1	Miroslav	Konecny	1978	7	10	Street	Saint John	49	Prague	412776	Czech
10	NM	Carla	Botni	1949	6	7	Avenue	Charles	null	Prague	412733	Czeck
7	M2	Anita	Rula	1982	9	7	Street	Sesto	23	Milano	20127	Italy
8	NM	Anna	Rla	null	null	null	Street	Sarca	336	Milano	20126	Italy
9	M3	Carlo	Batini	1949	6	7	Street	Beato Angelico	23	Milano	20133	Italy
2	NM	Martin	Necasky	1975	7	8	Square	Wien	null	Bratislava	101278	Slovakia
11	NM	Marta	Necasky	1976	11	8	null	null	null	Bratislava	112234	Slovakia

Scelta e acquisizione dati grezzi

Modellazione

Arricchimento semantico

Integrazione e fusione

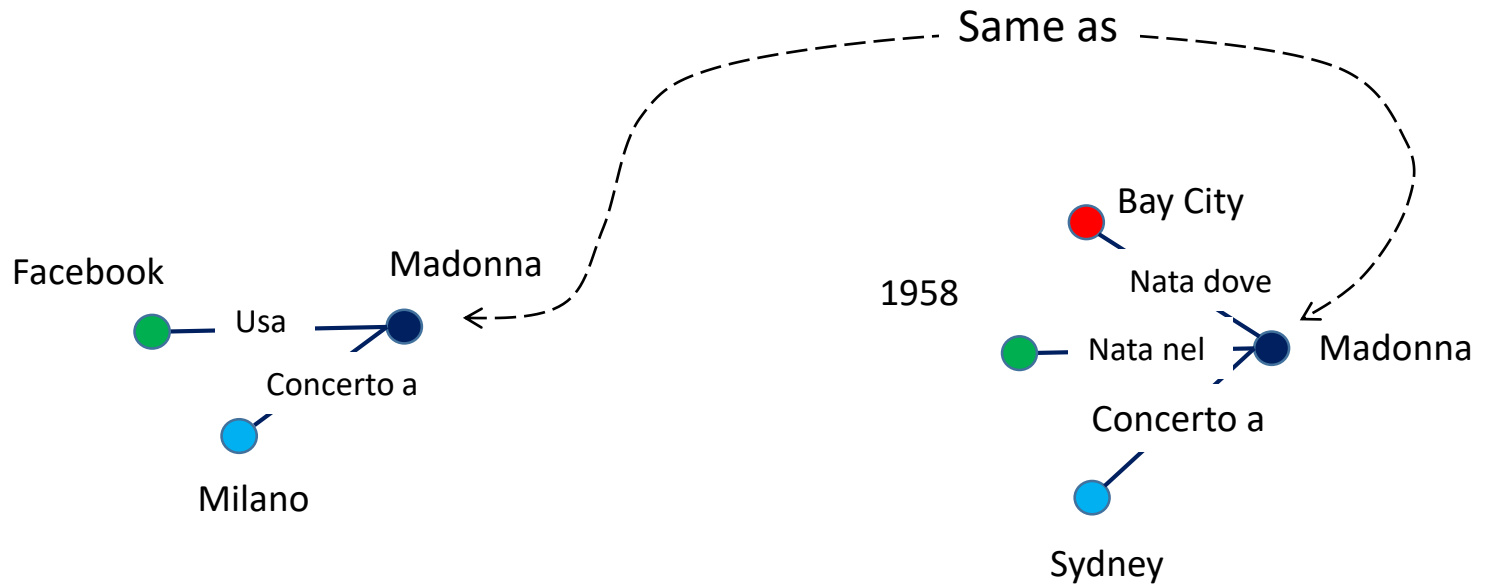
Condivisione

Analisi

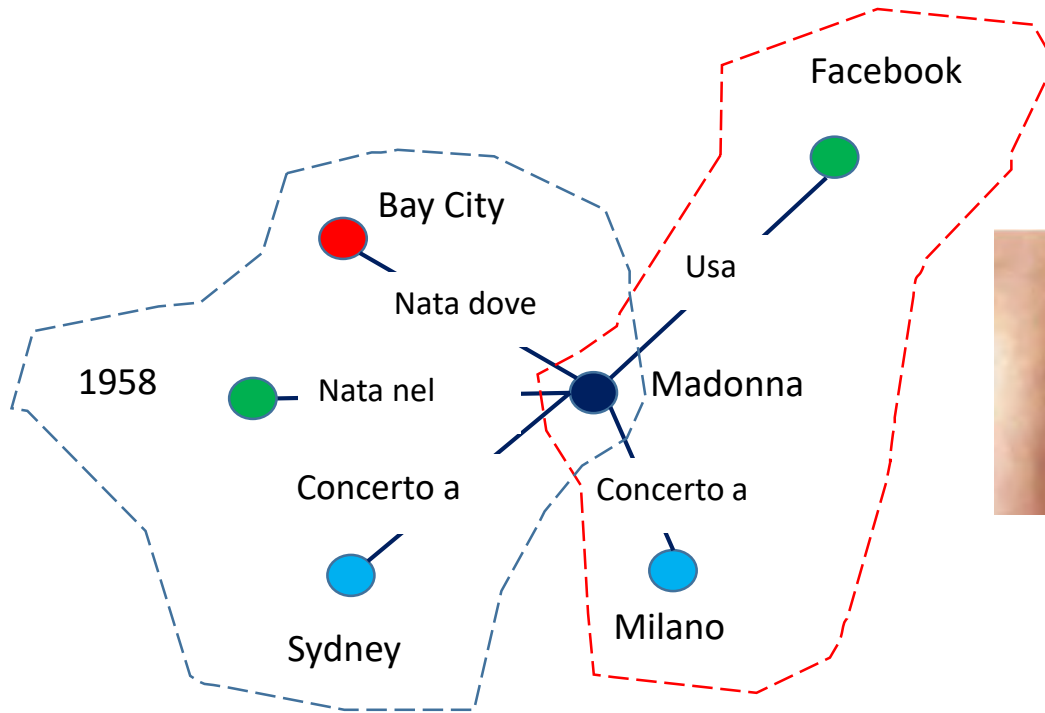
Visualizzazione

I dati vanno
condivisi

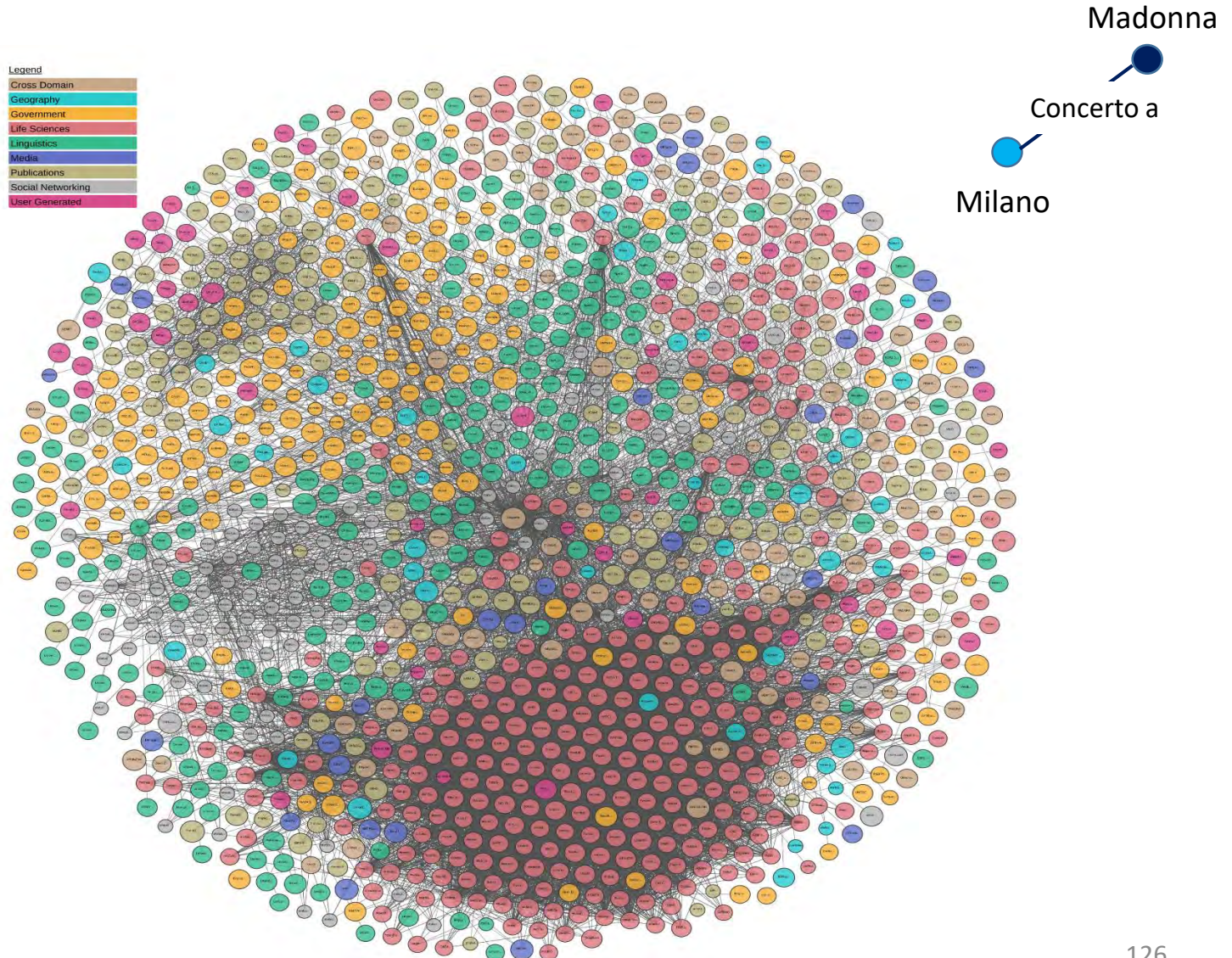
... Da Milano a Sydney



Milano + Sydney = Più Conoscenza condivisa (Win Win)



L'Open Data Cloud (100 miliardi di collegamenti)



Scelta e acquisizione dati grezzi

Modellazione

Arricchimento semantico

Integrazione e fusione

Condivisione

Analisi

Visualizzazione

Per risolvere i problemi
i dati vanno analizzati

Il problema della traduzione

Perchè non imparare dal Web?

La voce Ayrton Senna in Wikipedia

- Ayrton Senna da Silva (21 March 1960 – 1 May 1994) was a Brazilian racing driver who won the Formula One World Drivers' Championship in 1988, 1990 and 1991, and who is widely regarded as one of the greatest Formula One drivers of all time
- **Ayrton Senna da Silva** (San Paolo, 21 marzo 1960 – Bologna, 1 maggio 1994) è stato un pilota automobilistico brasiliano, campione del mondo di Formula 1 nel 1988, 1990 e 1991.
- Vincitore di 41 GP, è ritenuto uno dei piloti più forti di tutti i tempi.

Il problema della traduzione
Perchè non imparare dal Web?
La voce Ayrton Senna in Wikipedia

- **Ayrton Senna da Silva (21 March 1960 – 1 May 1994) was a Brazilian racing driver who won the Formula One World Drivers' Championship in 1988, 1990 and 1991, and who is widely regarded as one of the greatest Formula One drivers of all time**
- **Ayrton Senna da Silva** (San Paolo, **21 marzo 1960** – Bologna, **1 maggio 1994**) **è stato un pilota automobilistico brasiliano, campione del mondo di Formula 1 nel 1988, 1990 e 1991.**
- Vincitore di 41 GP, **è ritenuto uno dei piloti più forti di tutti i tempi.**



Esercizio – Trovare una voce su Wikipedia in cui la versione italiana sia tradotta dalla versione inglese

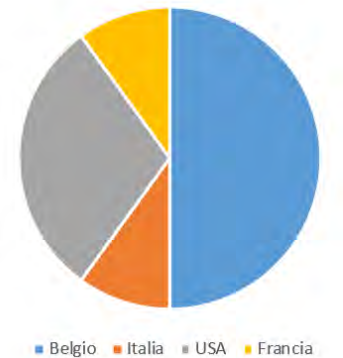
Tipologie di analisi

- Modelli descrittivi – Cosa accade, cosa è
- Modelli interpretativi – Perché accade
- Modelli predittivi – Cosa accadrà
- Modelli prescrittivi – Come intervenire sulla realtà, quale decisione prendere

Modelli descrittivi

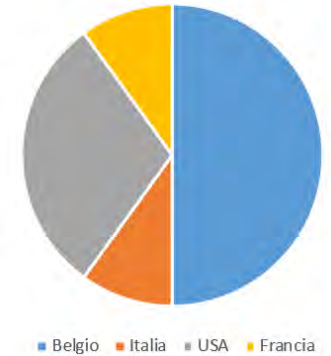
il caso dei professionisti europei

Miroslav	Konecny	1978	7	10	Street	St. John	49	Prague	412776	null
Martin	Necasky	1975	7	8	Sq.	Vienna	null	Bratislava	101278	Slovensko
Miroslav	Konecny	null	null	null	Str.	Saint Jon	49	Prague	412776	null
Crlo	Btini	1949	June	7	Street	Dessìè	15	Roma	00198	Italia
Miroslav	Knecy	1978	7	10	Sq.	Budapest	23	Wien	null	Austria
Anisa	Rula	1982	September	7	Via	Sesto	null	Milano	20...	Ital
Anita	Rula	1982	9	7	Via	Seto	23	Milan	null	Italy
Anna	Rla	null	null	null	Via	Sarca	336	Milano	null	Italy
Carlo	Batini	1949	6	7	V.	Beato Angelico	23	Milan	20133	Italy
Carla	Botni	1949	June	7	Av.	Charles	null	Prague	412733	null
Marta	Necasky	1976	11	8	null	null	null	Bratislava	112234	Slov.

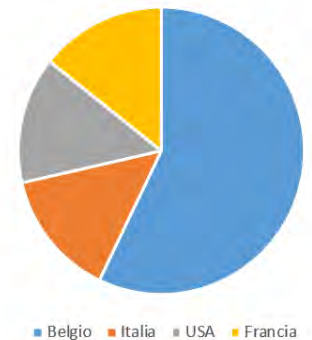


Dati di migliore qualità → Modelli di migliore qualità

Miroslav	Konecny	1978	7	10	Street	St. John	49	Prague	412776	null
Martin	Necasky	1975	7	8	Sq.	Vienna	null	Bratislava	101278	Slovensko
Miroslav	Konecny	null	null	null	Str.	Saint Jon	49	Prague	412776	null
Crlo	Btini	1949	June	7	Street	Dessiè	15	Roma	00198	Italia
Miroslav	Knecy	1978	7	10	Sq.	Budapest	23	Wien	null	Austria
Anisa	Rula	1982	September	7	Via	Sesto	null	Milano	20...	Ital
Anita	Rula	1982	9	7	Via	Seto	23	Milan	null	Italy
Anna	Rla	null	null	null	Via	Sarca	336	Milano	null	Italy
Carlo	Batini	1949	6	7	V.	Beato Angelico	23	Milan	20133	Italy
Carla	Botni	1949	June	7	Av.	Charles	null	Prague	412733	null
Marta	Necasky	1976	11	8	null	null	null	Bratislava	112234	Slov.



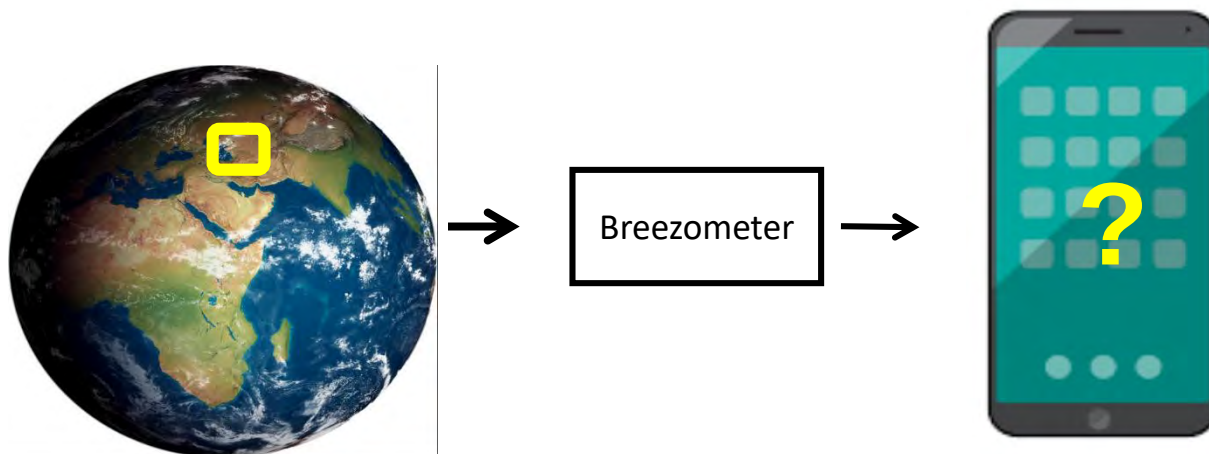
Tuple #	F.Name	Last Name	Y.Of Birth	M. Birth	D. Birth	Toponym	Name Toponym	Number T.	City	Zip Code	Country
1	Miroslav	Konecny	1978	7	10	Street	Saint John	49	Prague	412776	Czech
10	Carla	Botni	1949	6	7	Avenue	Charles	null	Prague	412733	Czech
7	Anita	Rula	1982	9	7	Street	Sesto	23	Milano	20127	Italy
8	Anna	Rla	null	null	null	Street	Sarca	336	Milano	20126	Italy
9	Carlo	Batini	1949	6	7	Street	Beato Angelico	23	Milano	20133	Italy
2	Martin	Necasky	1975	7	8	Square	Wien	null	Bratislava	101278	Slovakia
11	Marta	Necasky	1976	11	8	null	null	null	Bratislava	112234	Slovakia



Modelli predittivi

2. Modelli *predittivi*, che forniscono previsioni sul *futuro*, per esempio: *quali saranno i livelli di inquinamento un certo giorno in un certo luogo?*

- Questo è il caso della previsione dell'inquinamento



Modelli diagnostici e prescrittivi

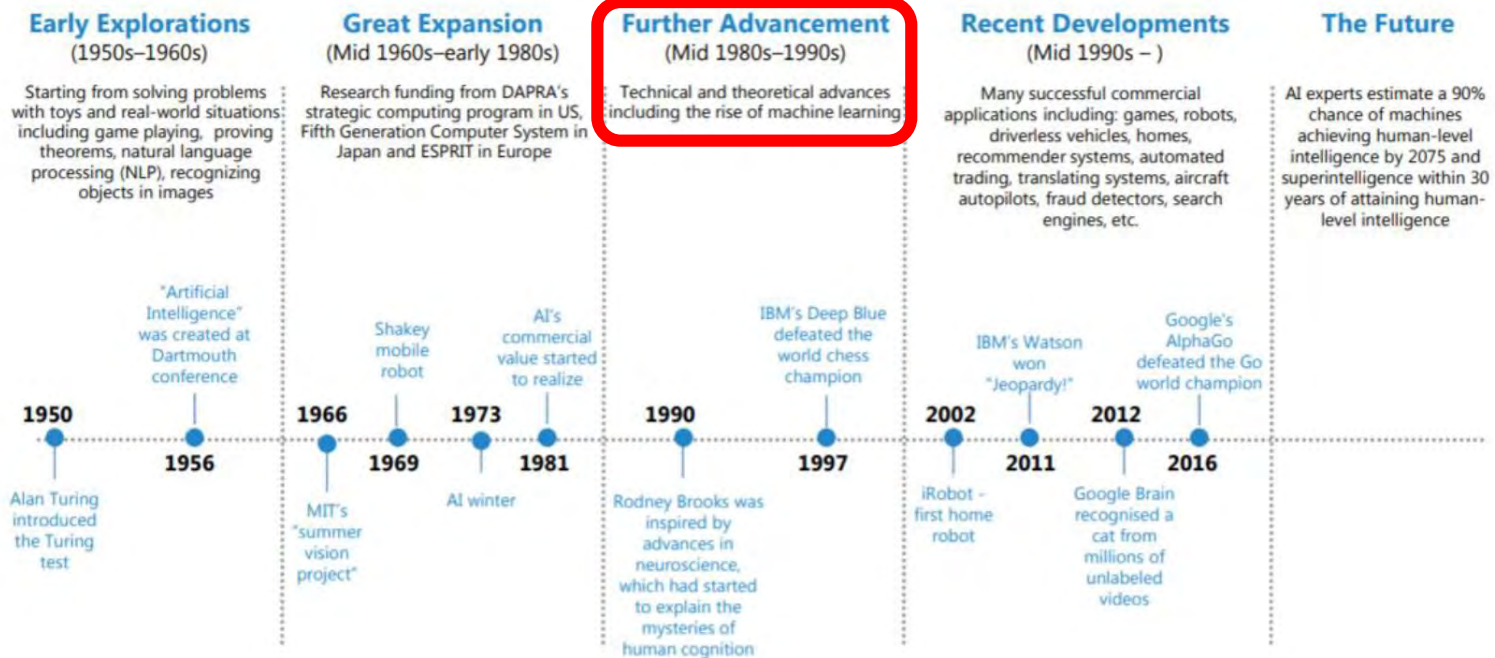
3. Modelli *diagnostici*, o *interpretativi*, che forniscono una risposta alla domanda: *perché* è accaduto un certo evento, come, ad esempio, *perché un certo luogo è molto inquinato?*

4. Modelli *prescrittivi*, che forniscono strategie e decisioni che ci permettono di far accadere ciò che vogliamo, per esempio *ottimizzare dinamicamente il numero di posti disponibili nelle diverse corse di treni AV nell'anno.*

Torniamo alla Intelligenza Artificiale
che avevamo lasciato moribonda...

Evoluzione della IA negli anni la ripresa di interesse

AI Evolution Over the Years



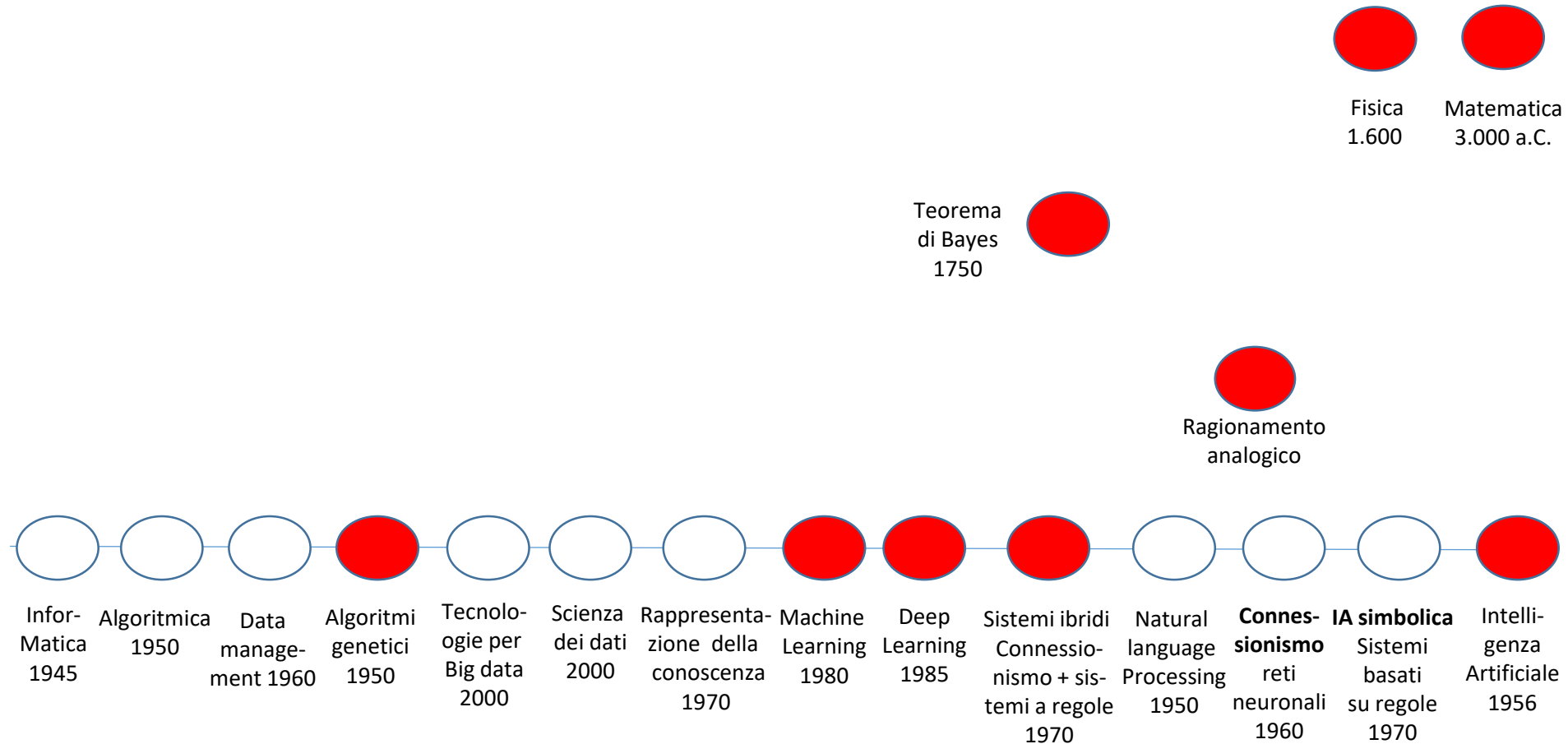
<https://www.slideshare.net/yanaioron1/vertex-perspectives-artificial-intelligence>

La grande svolta della nuova IA

Si passa da

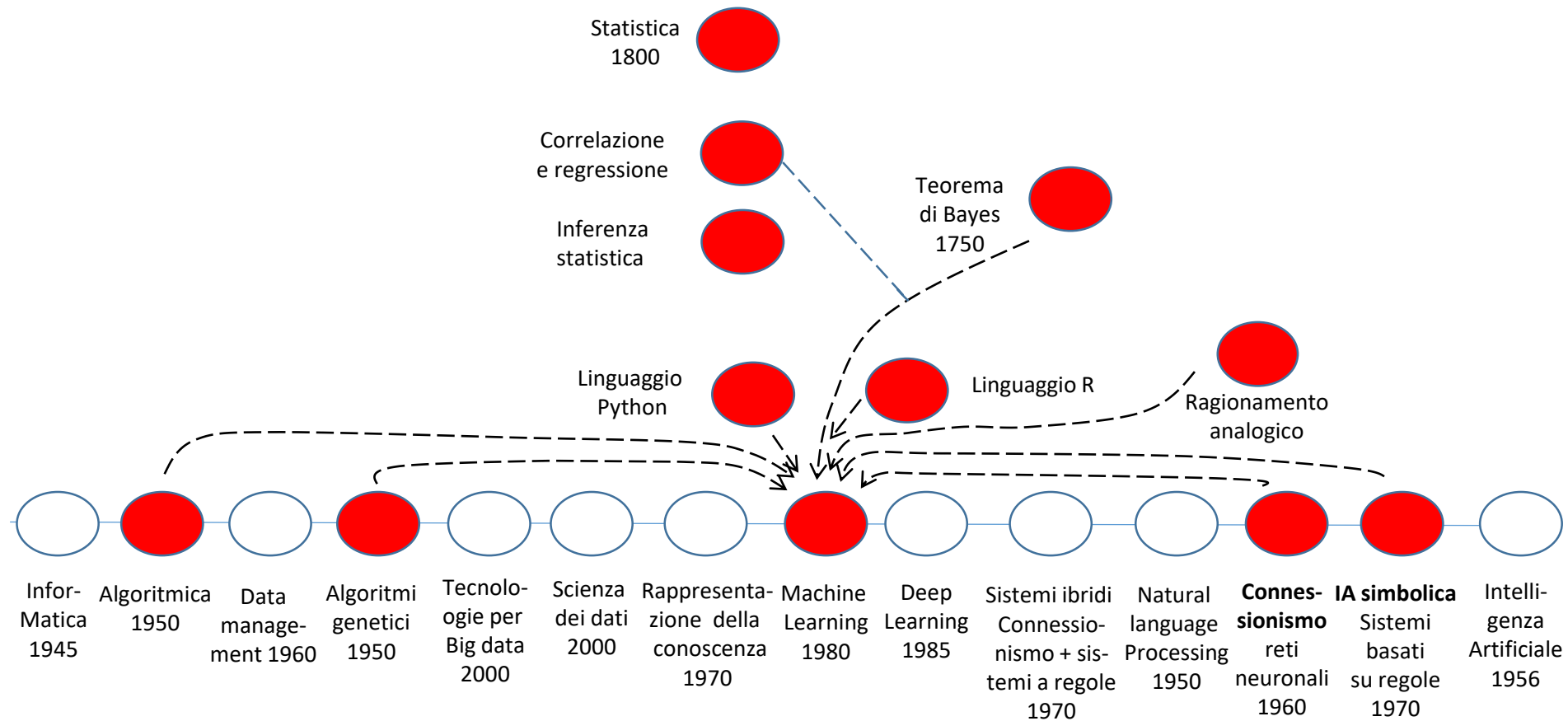
- cercare di imitare il **funzionamento della mente/cervello umano** a
- imitare il **modo con cui gli umani apprendono**

La nuova IA



II Machine Learning

Paradigmi del Machine learning [Domingos 2015]



Paradigmi del Machine Learning

- **Simbolico** – L'apprendimento consiste nel saper manipolare simboli logici
- **Connessionistico** – L'apprendimento è ciò che calcola il cervello con i suoi 10 miliardi di neuroni connessi tra loro
- **Evoluzionistico** – L'apprendimento è la selezione naturale
- **Bayesiano** – L'apprendimento è la evoluzione della incertezza governata dall'inferenza.
- **Analogico** – L'apprendimento consiste nel trovare similarità tra fenomeni, così da inferire nuove similarità

Tipologie di machine learning (apprendimento automatico)

- ML supervisionato – il ML apprende da esempi
- ML non supervisionato – il ML riconosce strutture e pattern senza avere a disposizione esempi
- ML di reinforcement – il ML è ispirato da feedback positivo (premio) o negativo (multa)

Apprendimento supervisionato

Previsione inquinamento

Apprendimento supervisionato: all'algoritmo di apprendimento sono forniti, come **esperienza**, dati di **input** (**tempo, traffico, ecc.**) e rispettivi dati di **output** (**livelli di inquinamento**).

Questi dati vengono suddivisi in **dati di training** su cui l'algoritmo **apprende**, e **dati di test**, a cui l'algoritmo è applicato per **testare la sua qualità**.

Input

Output

latitudine	longitudine	Anno	Mese	Giorno	Ora	Minuto	Traffico	Riscalda- mento	Qualità dell'aria	Livello di Inquina- mento
308.40	215.20	2017	6	12	8	10	0.7	0.7	47	3
308.40	215.20	2017	7	12	8	20	0.7	0.8	49	2
308.40	215.20	2017	6	12	8	30	0.8	0.8	51	5
308.40	215.20	2017	6	12	8	40	0.9	0.7	54	6
308.40	215.20	2017	6	12	8	50	0.8	0.6	48	4

**Dati di
training**

**Dati di
test**

Metodi basati su learning apprendimento supervisionato

A questo punto, quando vengono forniti dati per una nuova previsione (es. oggi pomeriggio a Piazza Duomo a Milano) l'algoritmo di apprendimento supervisionato applica il modello costruito con i dati di training ai nuovi dati

latitudine	longitudine	Anno	Mese	Giorno	Ora	Minuto	Traffico	Riscaldamento	Qualità dell'aria	Livello di inquinamento
308.40	215.20	2017	6	12	8	10	0.7	0.7	47	3
308.40	215.20	2017	7	12	8	20	0.7	0.8	49	2
308.40	215.20	2017	6	12	8	30	0.8	0.8	51	5
308.40	215.20	2017	6	12	8	40	0.9	0.7	54	6
..	..	..	..	..	..	.			..	4
308.45	215.25	2020	02	10	18	10	0.7	0.7	54	
308.45	215.25	2020	02	10	18	20	0.7	0.7	57	
308.45	215.25	2020	02	10	18	30	0.8	0.8	60	
308.45	215.25	2020	02	10	18	40	0.9	0.7	65	
308.45	215.25	2020	02	10	18	50	0.8	0.6	70	

Metodi basati su learning - apprendimento supervisionato

A questo punto, quando vengono forniti dati per una nuova previsione (es. oggi pomeriggio a Piazza Duomo a Milano) l'algoritmo di apprendimento supervisionato applica il modello costruito con i dati di training ai nuovi dati

latitudine	longitudine	Anno	Mese	Giorno	Ora	Minuto	Traffico	Riscaldamento	Qualità dell'aria	Livello di inquinamento
308.40	215.20	2017	6	12	8	10	0.7	0.7	47	3
308.40	215.20	2017	7	12	8	20	0.7	0.8	49	2
308.40	215.20	2017	6	12	8	30	0.8	0.8	51	5
308.40	215.20	2017	6	12	8	40	0.9	0.7	54	6
..	..	..	..	..	..	.			..	4
308.45	215.25	2020	02	10	18	10	0.7	0.7	54	3
308.45	215.25	2020	02	10	18	20	0.7	0.7	57	2
308.45	215.25	2020	02	10	18	30	0.8	0.8	60	5
308.45	215.25	2020	02	10	18	40	0.9	0.7	65	2
308.45	215.25	2020	02	10	18	50	0.8	0.6	70	5

Apprendimento non supervisionato

All'algoritmo di apprendimento sono forniti, come esperienza, dati di input senza che questi siano associati a valori di output, ed esso ha il compito di riconoscere schemi/strutture nell'insieme dei dati fornito.

Rientrano nella classe degli algoritmi di apprendimento non supervisionato quelli di **clustering**, ossia quegli algoritmi che hanno il compito di raggruppare i dati sulla base di strutture, schemi o affinità che essi devono riconoscere.

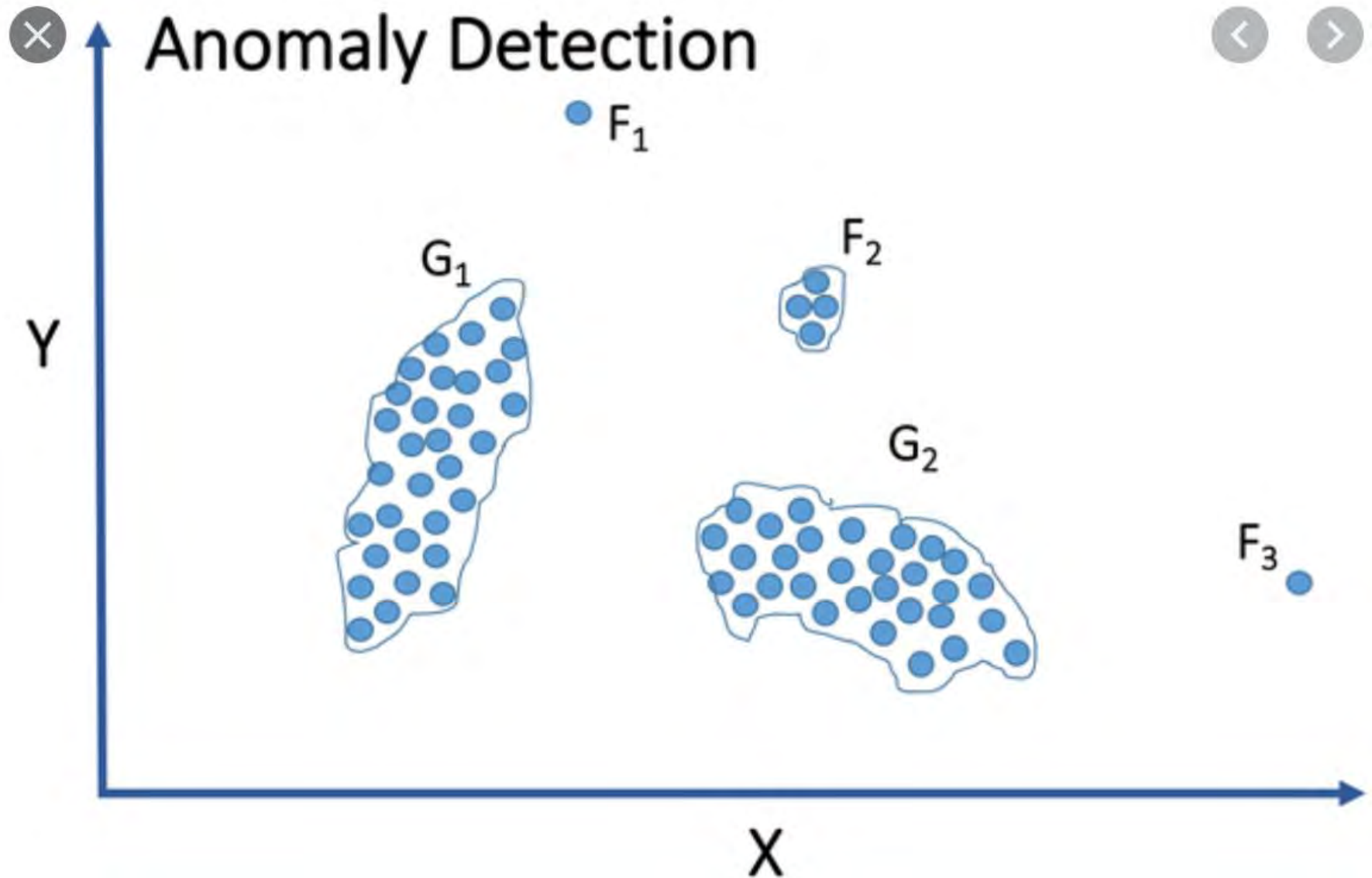


Saggistica



Narrativa

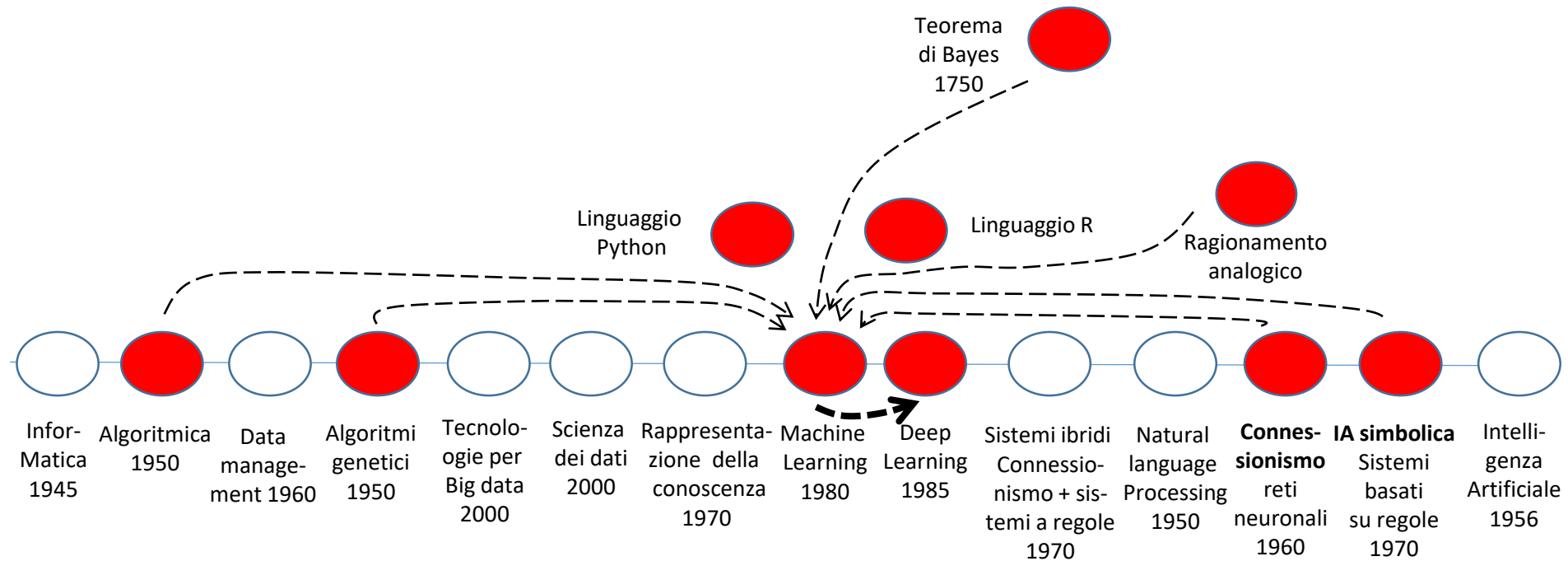
Anomaly detection mediante clustering



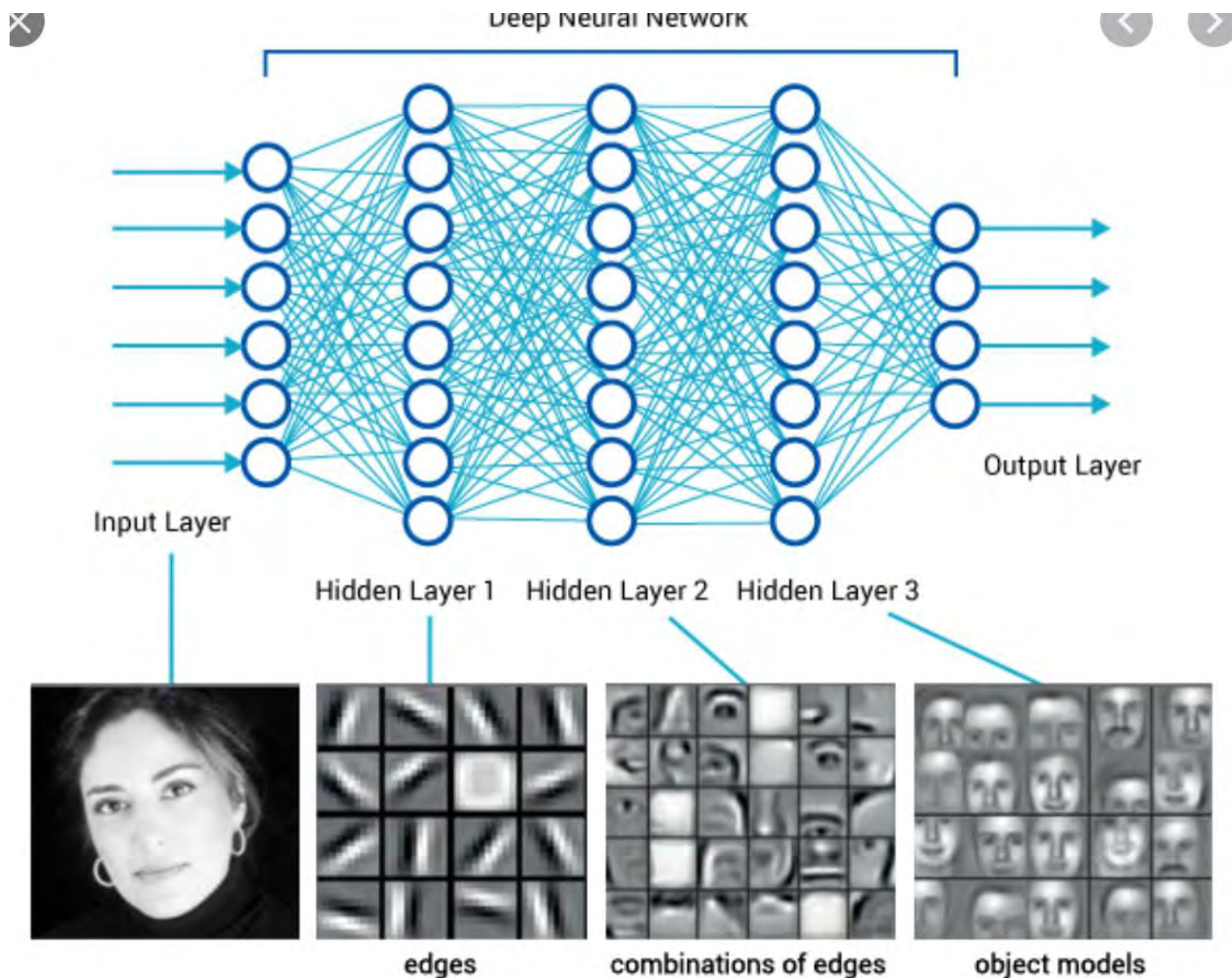
Apprendimento per rinforzo

- L'algoritmo di apprendimento apprende **sulla base della risposta dell'ambiente alle sue azioni**; in tal caso, infatti, vi è un'interazione con un ambiente nel quale esso deve svolgere un compito e dal quale ottiene una risposta in termini di “**ricompensa**” (il rinforzo) che consiste nella valutazione della prestazione.
- Come esempi, citiamo gli algoritmi che imparano le strategie da usare in un gioco attraverso partite contro un avversario e quelli che consentono ad un robot di muoversi all'interno di un ambiente con un certo scopo.

Il Deep Learning



Le reti convoluzionali nel riconoscimento facciale



Scelta e acquisizione dati grezzi

Modellazione

Arricchimento semantico

Integrazione e fusione

Condivisione

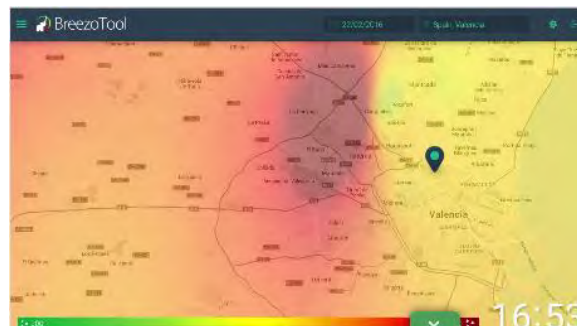
Analisi

Visualizzazione

I dati vanno visualizzati

Una heat map per l'inquinamento

latitudine	longitudine	Anno	Mese	Giorno	Ora	Minuto	Traffico	Riscaldamento	Qualità dell'aria	Livello di inquinamento
308.40	215.20	2017	6	12	8	10	0.7	0.7	47	3
308.40	215.20	2017	6	12	8	20	0.7	0.8	49	2
308.40	215.20	2017	6	12	8	30	0.8	0.8	51	5
308.40	215.20	2017	6	12	8	40	0.9	0.7	54	6
..	..	..	..	..	..	.			..	4
308.45	215.25	2017	6	12	8	10	0.7	0.7	54	5
308.45	215.25	2017	6	12	8	20	0.7	0.7	57	6
308.45	215.25	2017	6	12	8	30	0.8	0.8	60	9
308.45	215.25	2017	6	12	8	40	0.9	0.7	65	6
308.45	215.25	2017	6	12	8	50	0.8	0.6	70	4

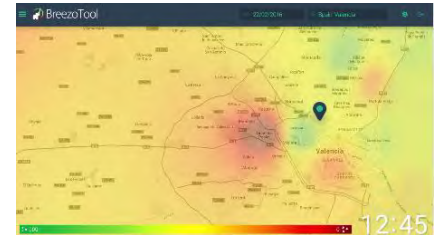


L'esempio Breezometer

Date due coordinate di latitudine e longitudine, quale sarà il livello di inquinamento domani nel punto del mondo che corrisponde a queste due coordinate?

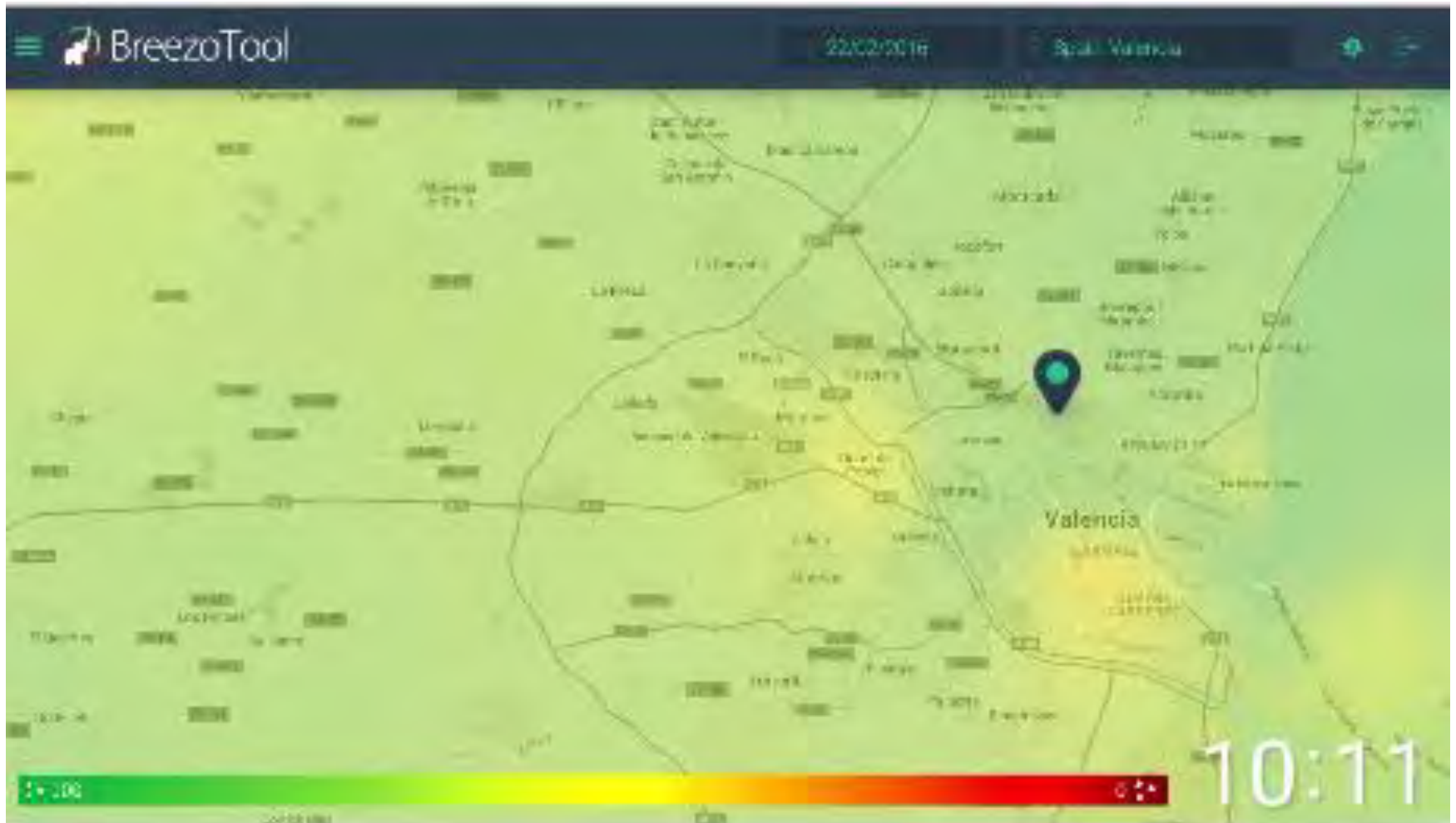


Breezometer

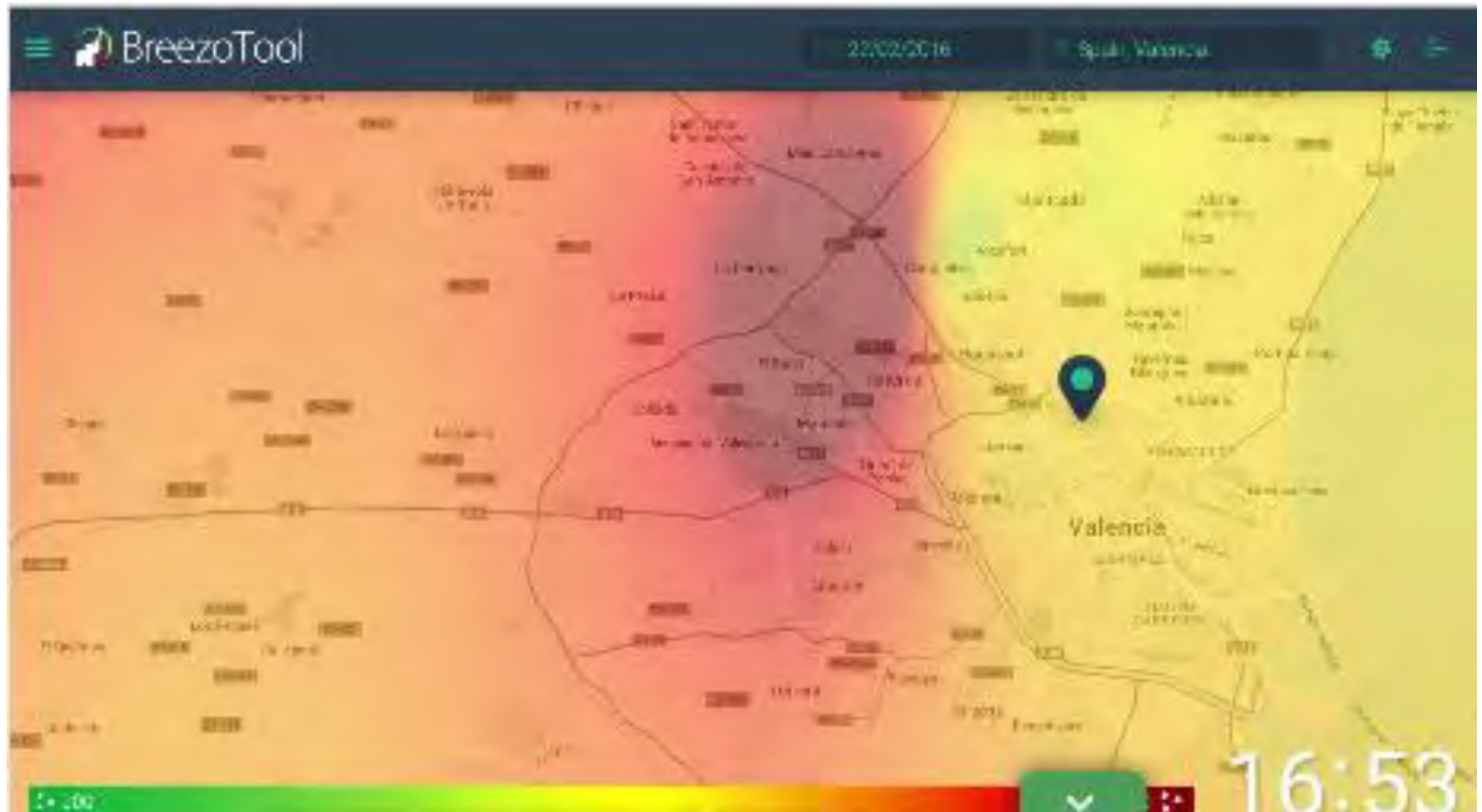


Livello di inquinamento in un territorio in un dato momento della giornata

Visualizzazione in Breezometer ore 10.11



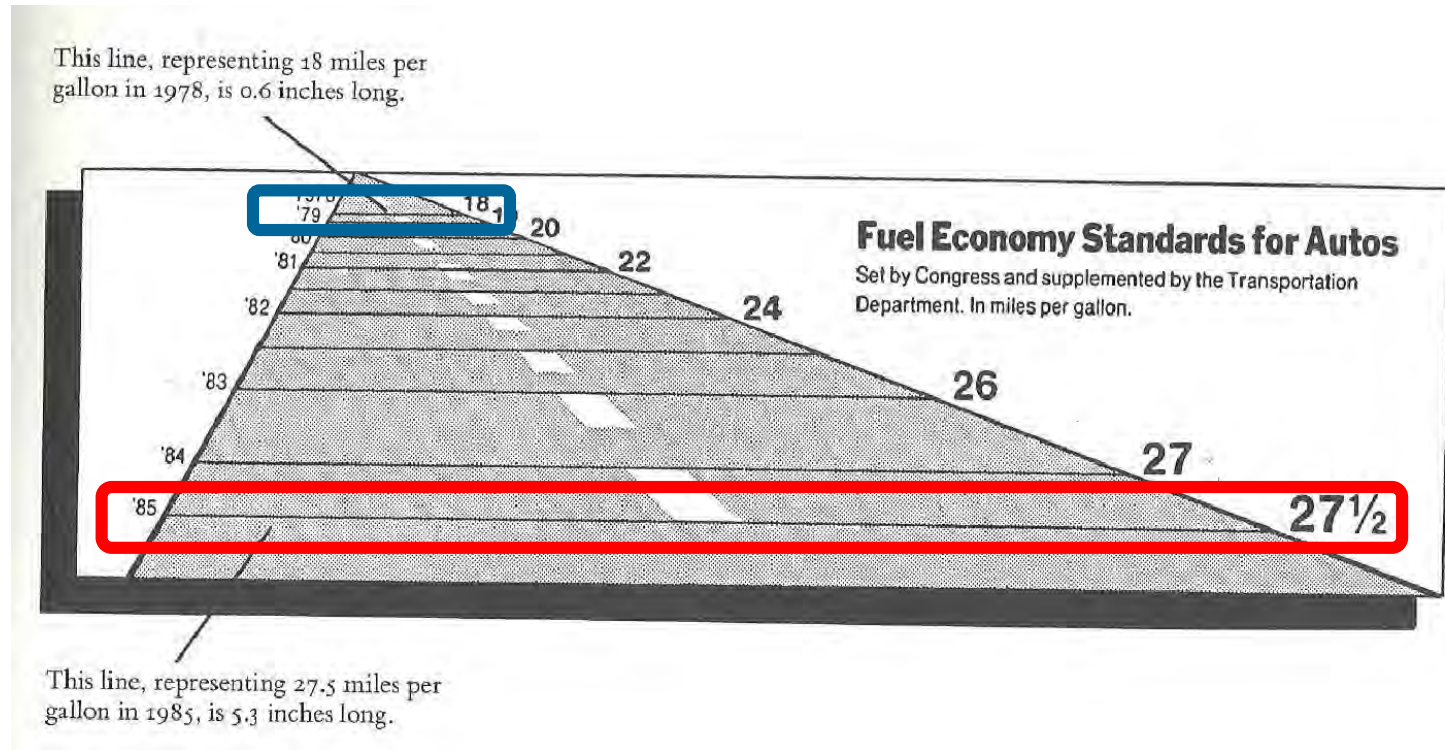
Ore 16.53



Attenzione!

Quante bugie nelle visualizzazioni...

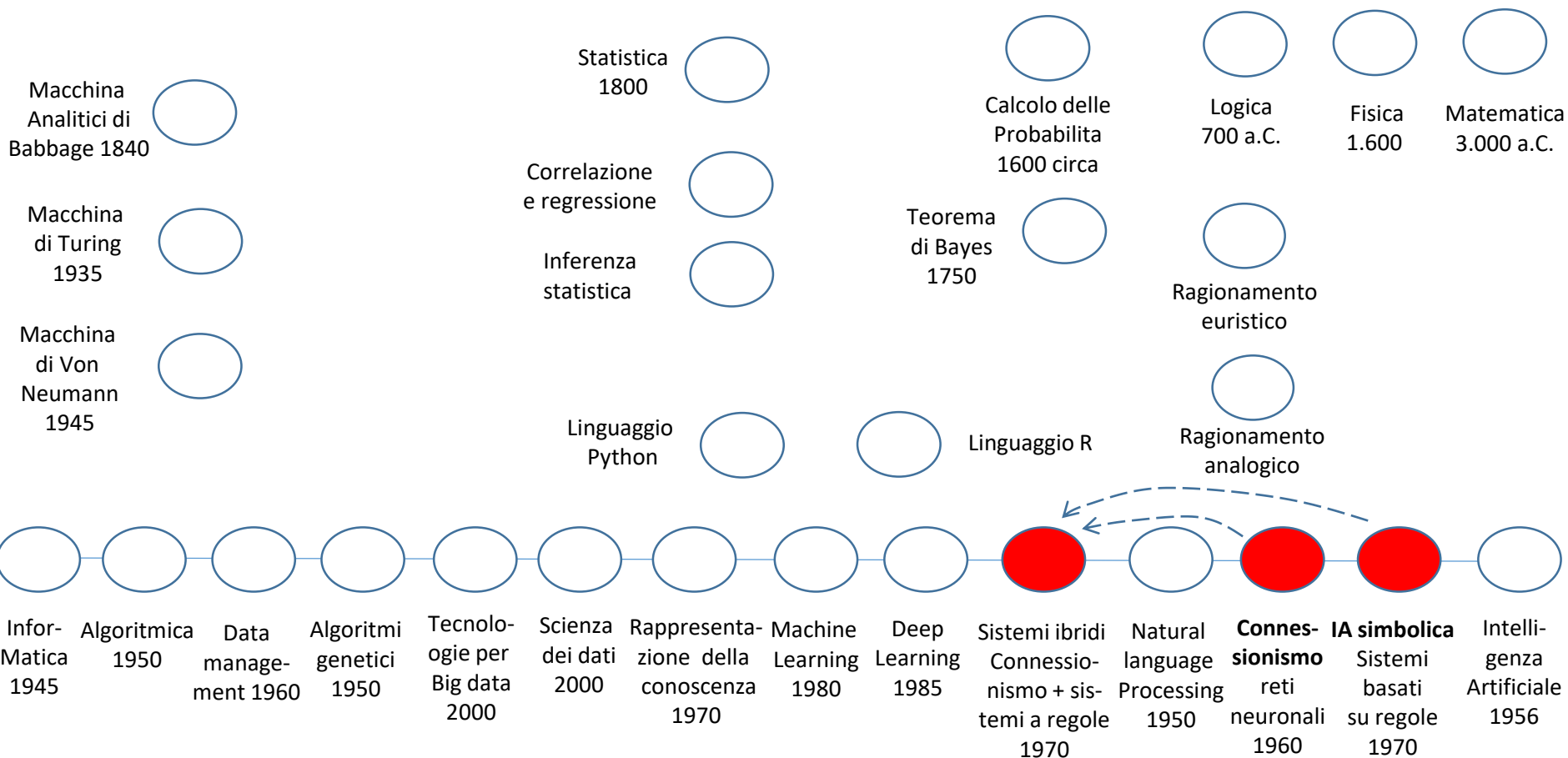
Anno	Miles per gallon
1978	18
1979	19
1980	20
1981	22
1982	24
1983	26
1984	27
1985	27,5



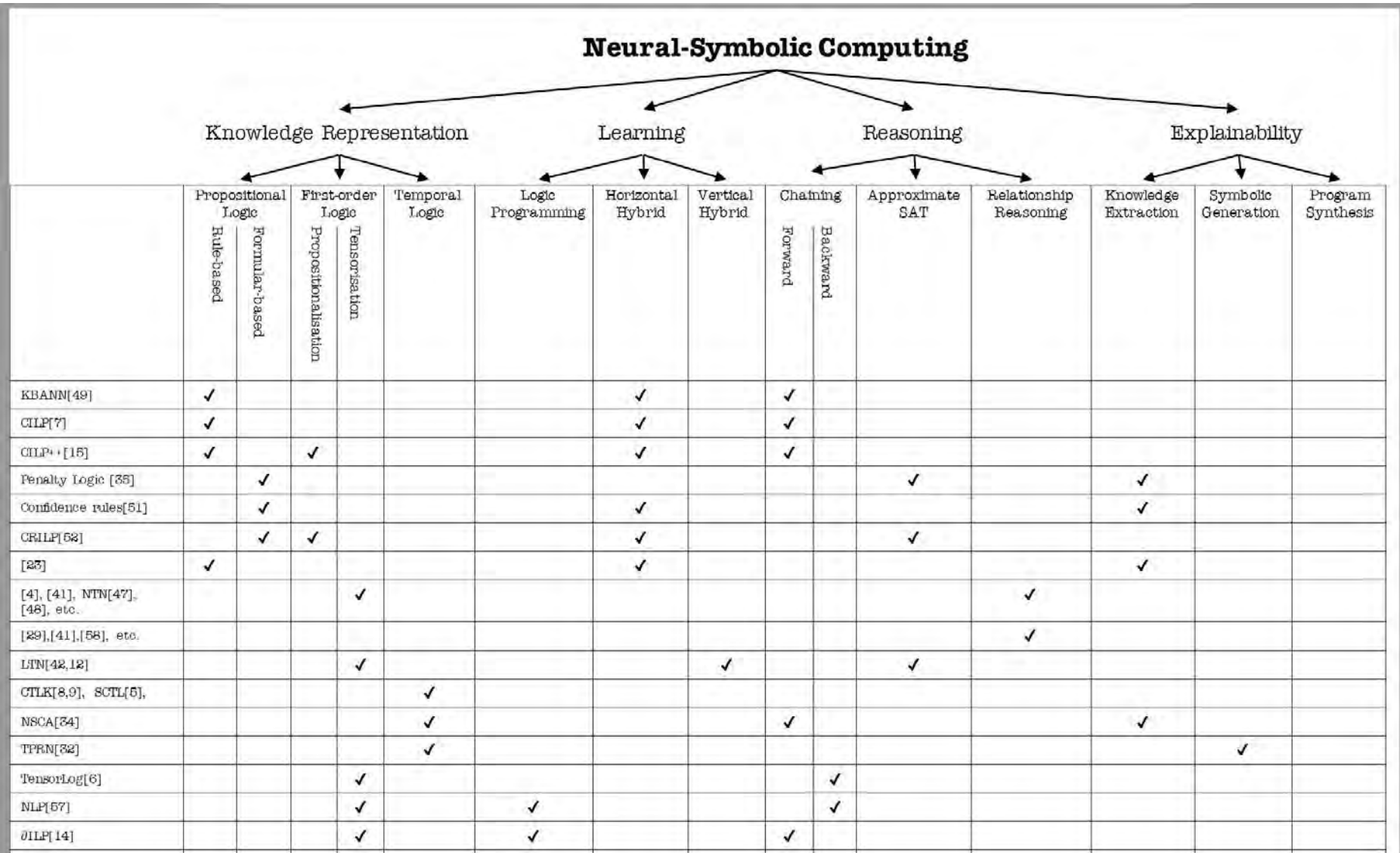
Livello di «bugia» = rapporto tra lunghezze nella visualizzazione / rapporto tra valori numerici nel mondo reale = 15

Gli sviluppi più recenti

I sistemi ibridi tra IA simbolica e connessionistica il Neural Symbolic Computing



Neural Symbolic Computing



Neural-Symbolic Computing: An Effective Methodology for Principled Integration of Machine Learning and Reasoning – Maggio 2019

Neural-symbolic computing aims at integrating two most fundamental cognitive abilities: the ability to learn from the environment, and the ability to reason from what has been learned.

Neural-symbolic computing has been an active topic of research for many years, reconciling the advantages of robust learning in neural networks and reasoning and interpretability of symbolic representation.

In this paper, we survey recent accomplishments of neural-symbolic computing as a principled methodology for integrated machine learning and reasoning.

We illustrate the effectiveness of the approach by outlining the main characteristics of the methodology: principled integration of neural learning with symbolic knowledge representation and reasoning allowing for the construction of explainable AI systems.

The insights provided by neural-symbolic computing shed new light on the increasingly prominent need for interpretable and accountable AI systems

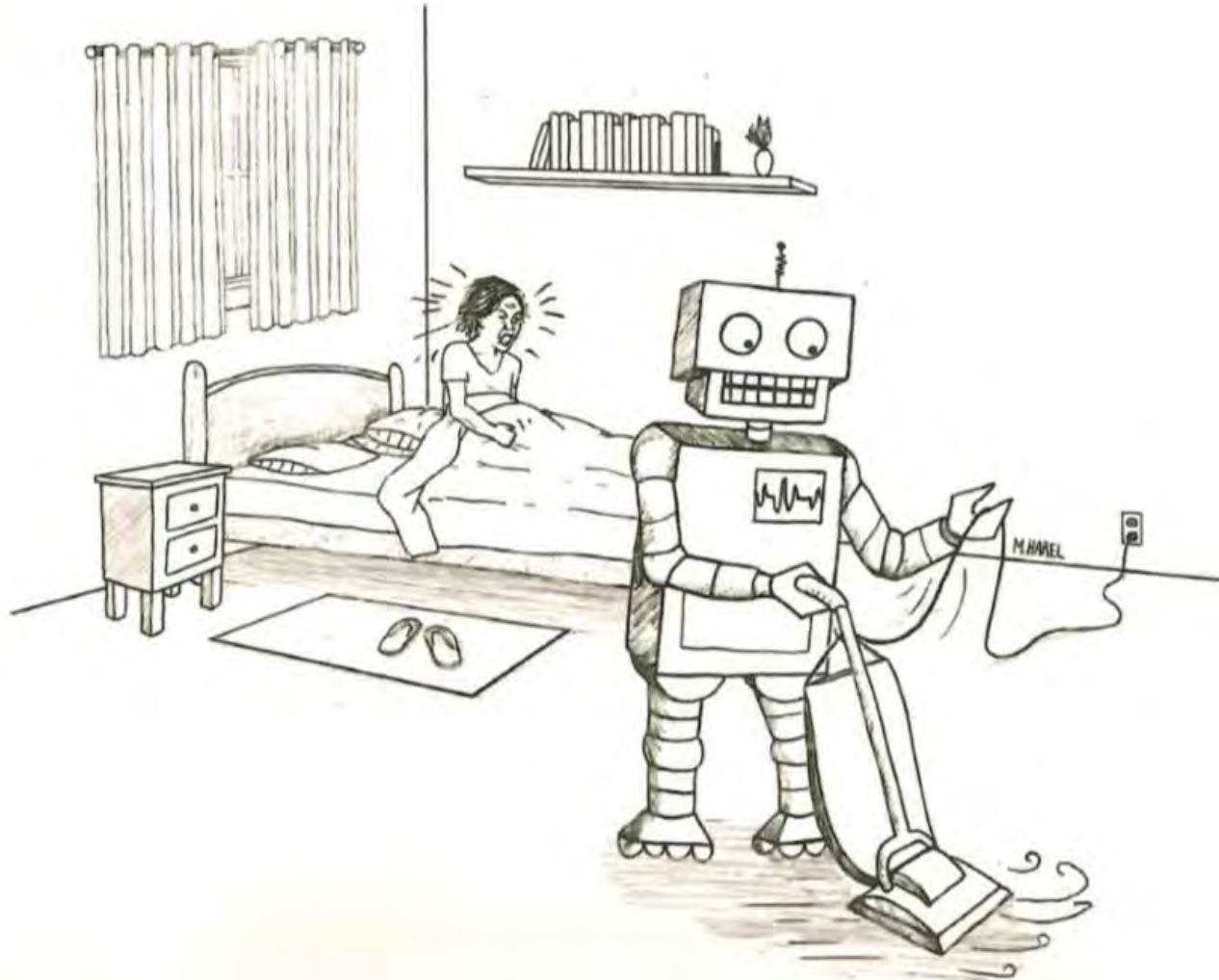
Il limite fondamentale del machine learning data driven - 1

D

ata is old news. By its nature, it tells us about things that have already happened. So to make decisions about the future, industries from engineering to bioscience still turn to simulation.

Wired, May 2017

Il robot intelligente che inizia a lavorare alle 6 ...



Il limite fondamentale del machine learning data driven - 2

LIMITATIONS OF BIG DATA

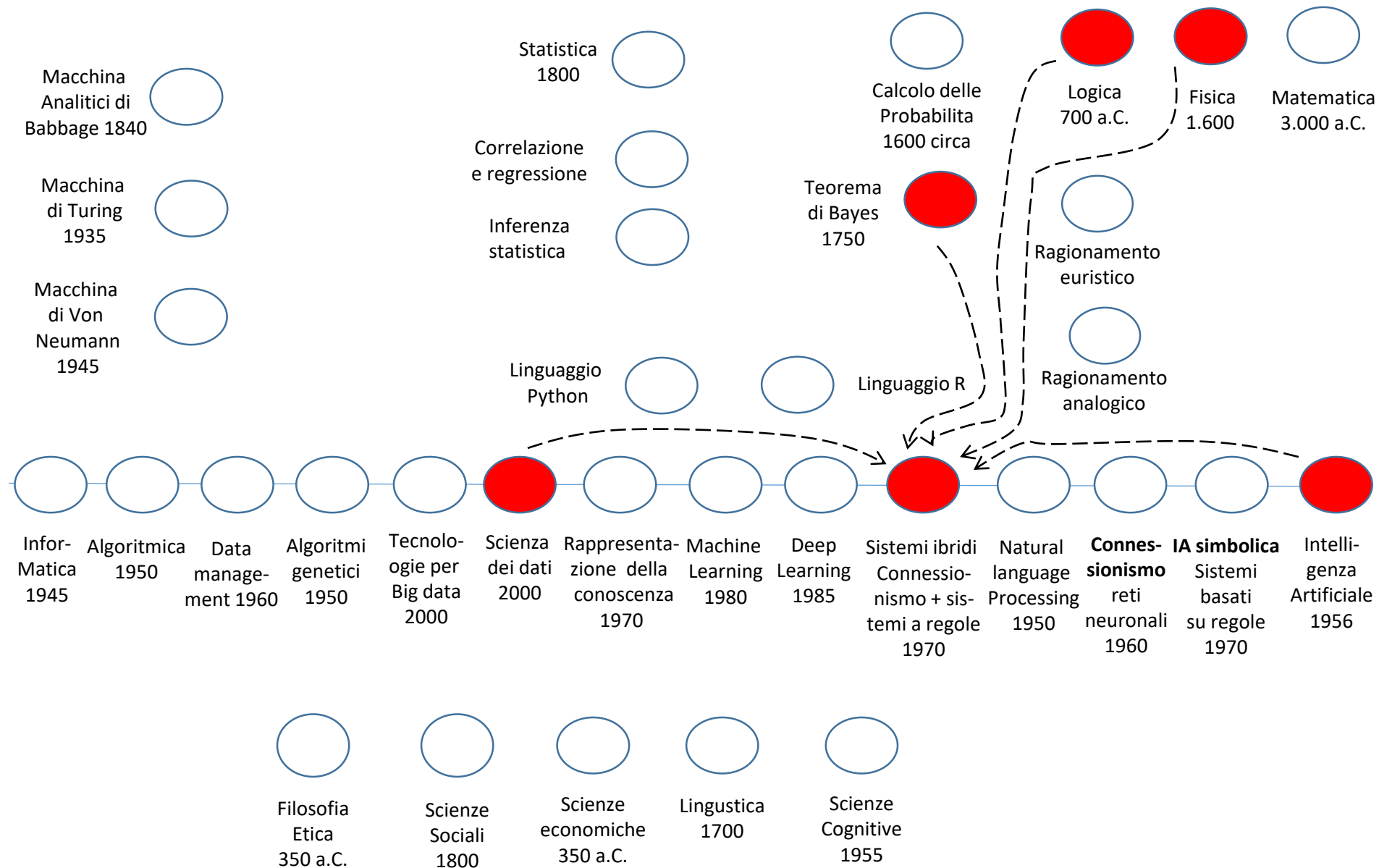
- In many areas, large data sets are non existent
 - What do aliens look like?
 - If every vehicle in San Francisco was self-driving, how would we redesign traffic lights?
 - How to take subways in Tokyo based on experiences in London or New York?

Il limite fondamentale del machine learning data driven - 3

WHAT-IF QUESTIONS THAT GOOGLE CANNOT ANSWER

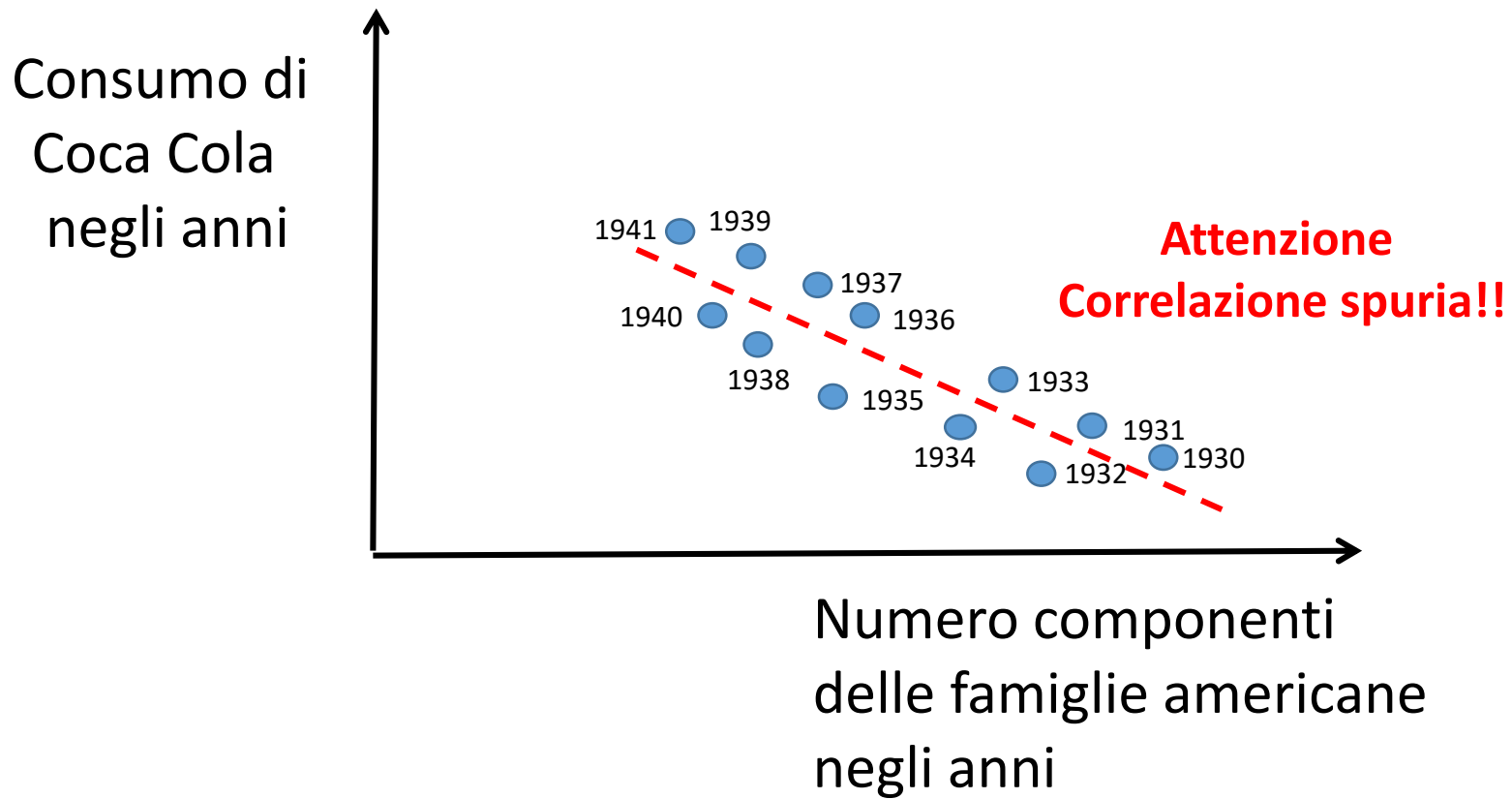
- If gas (petrol) was free, would more or fewer people buy electric cars?
- If the city of Chicago paid \$1 000 for every gun that is turned in, would that reduce gun violence?
- If humans had wings, would a wall between the US and Mexico be effective?

La nuova scienza della causa e dell'effetto



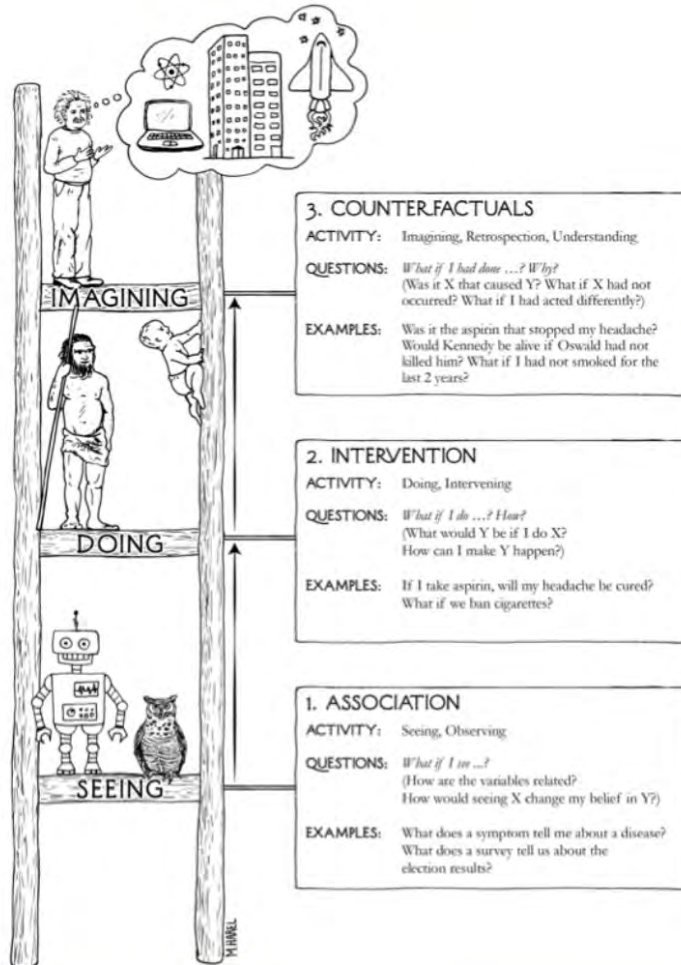
Correlazione

Consumo di Coca Cola e numero di componenti delle famiglie americane



Judea Pearl – Il libro del perchè

Ladder of Causation



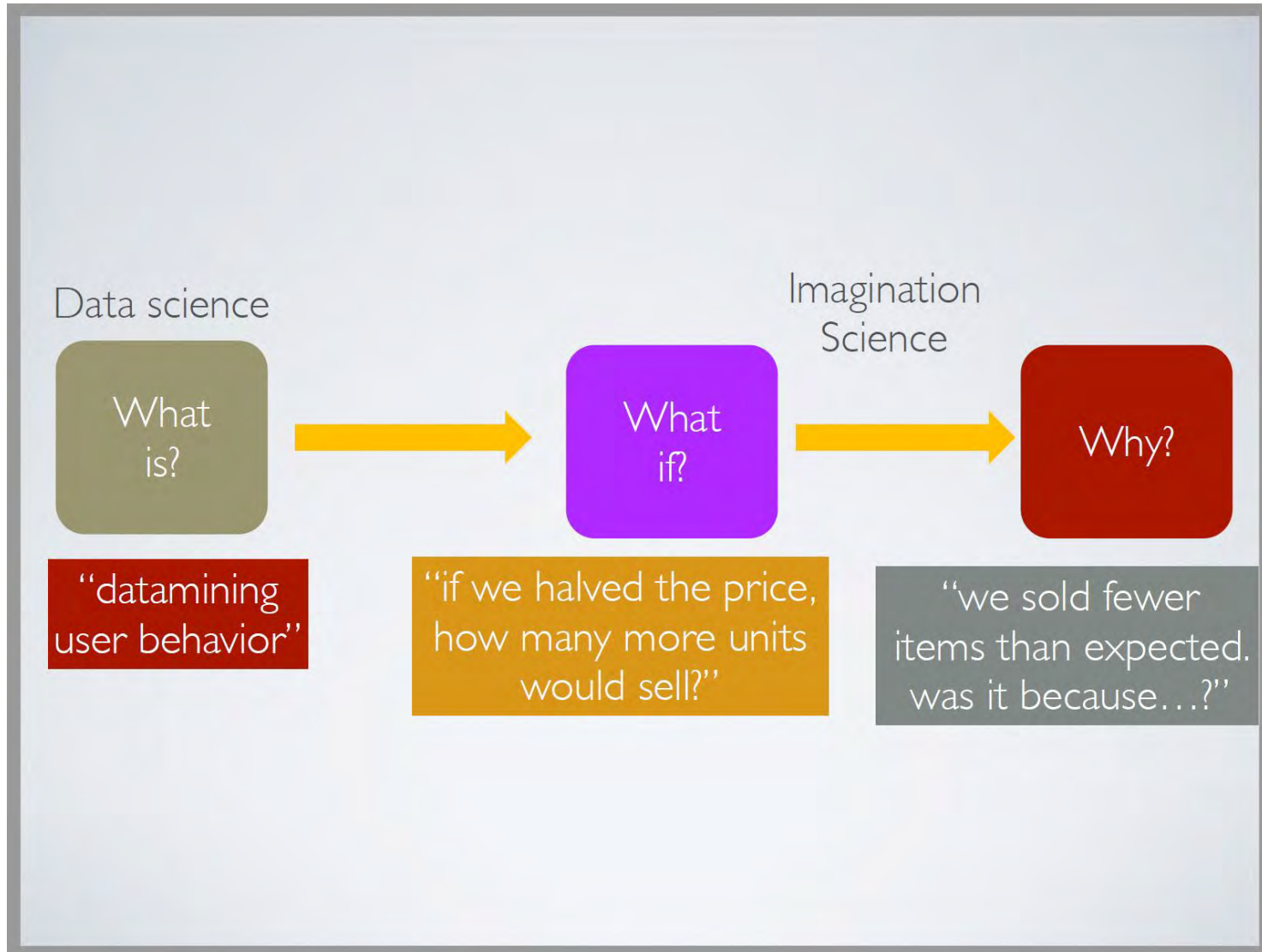
JUDEA PEARL
WINNER OF THE TURING AWARD
AND DANA MACKENZIE

THE
BOOK OF
WHY

α  β

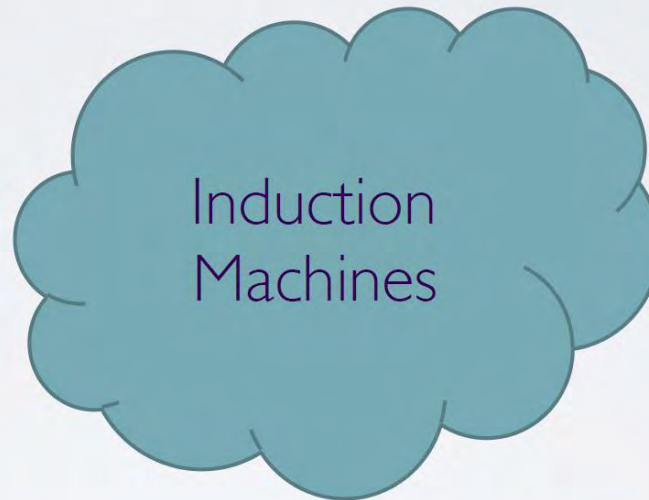
THE NEW SCIENCE
OF CAUSE AND EFFECT

Cosa è? Cosa accadrebbe se? Perché?



Livello 1 - Association

Estimate $P(Y = \text{"Scrub Jay"} \mid X = \text{image})$
Learning from **observations**



"Scrub Jay"

Most of modern machine learning
(e.g., deep learning) is at level 1

Livello 2 – Intervention

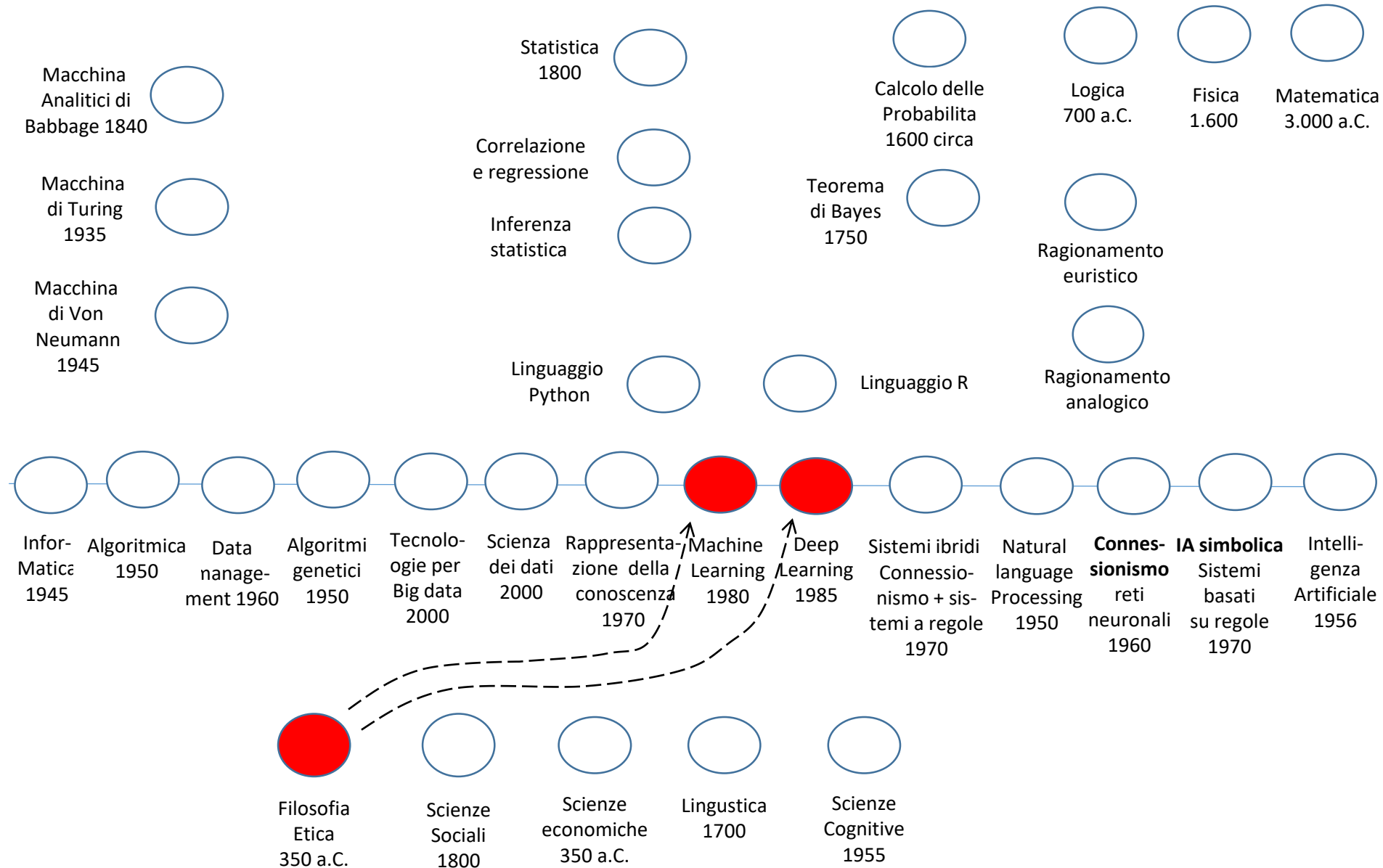
Fundamental difference between observation and intervention

$P(y \mid x, a)$ vs. $P(y \mid x, \text{do}(a))$

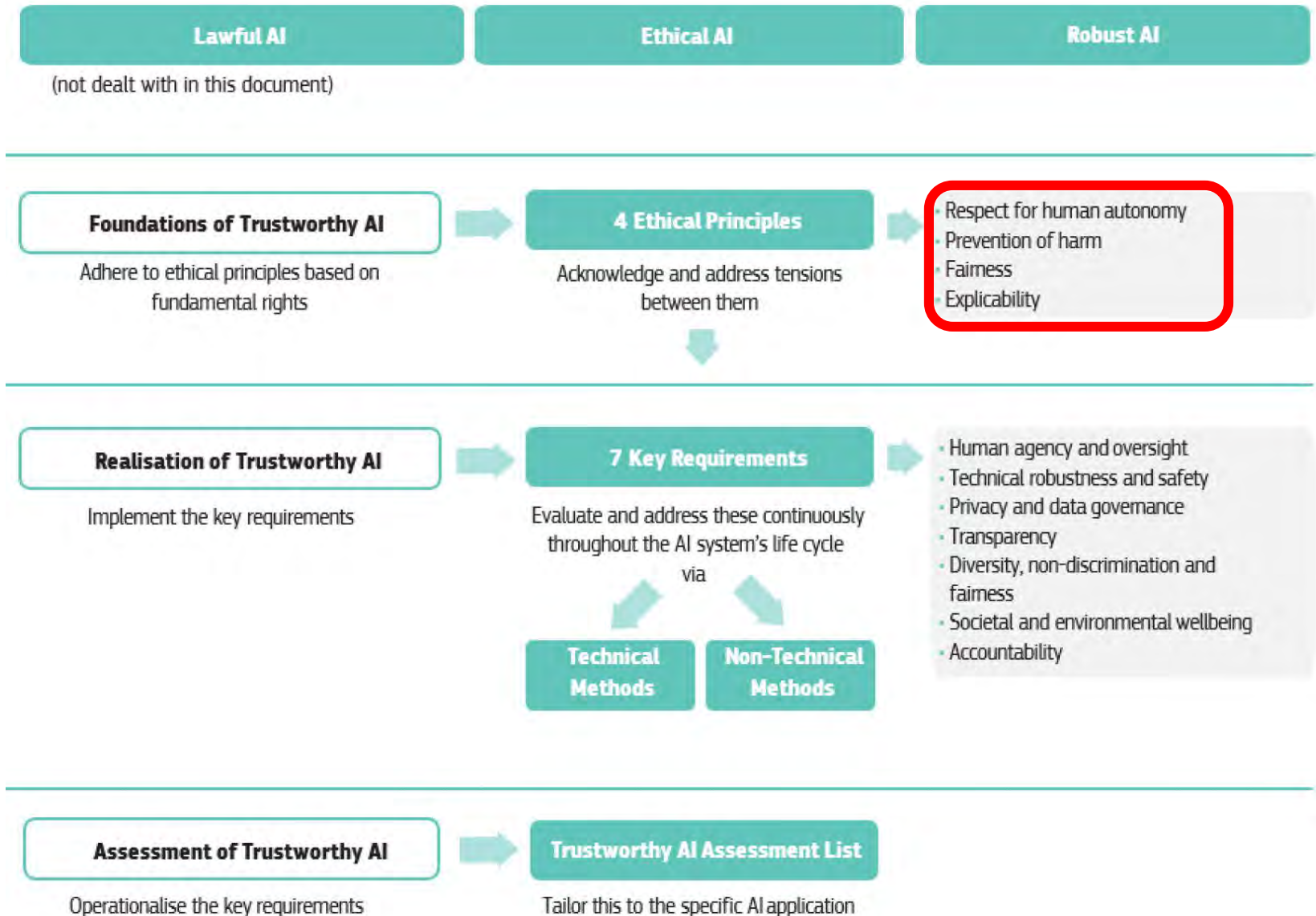
$P(Y = \text{"Hummingbird"} \mid X, \text{do}(A = \text{"install Hummingbird feeder"}))$



Etica e Learning



Per una IA di cui ci si può fidare



Beyond *State v. Loomis*: Artificial Intelligence, Government Algorithmization, and Accountability

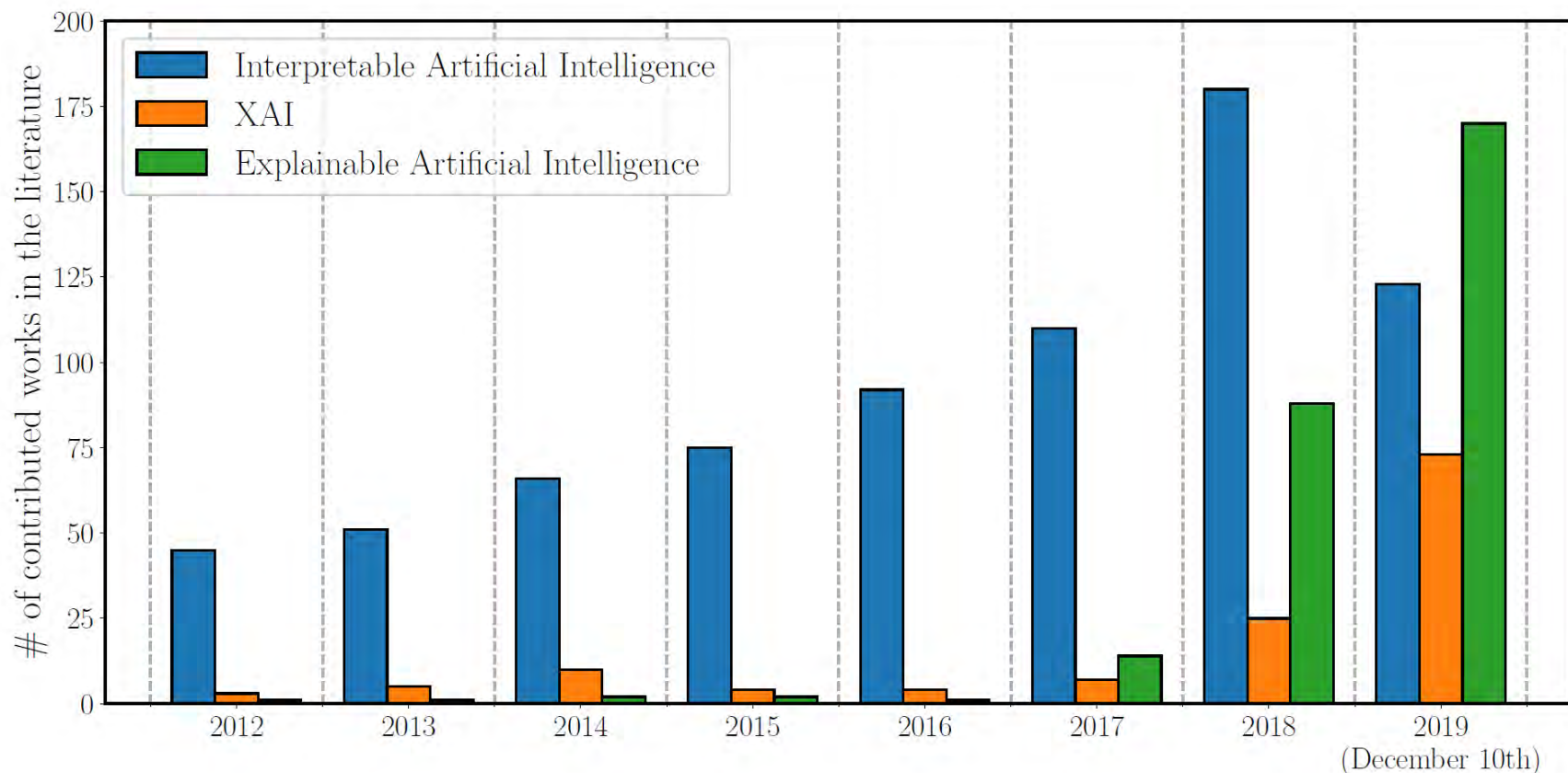
Developments in data analytics, computational power, and machine learning techniques have driven all branches of the government to outsource authority to machines in performing public functions—social welfare, law enforcement, and most importantly, courts. Complex statistical algorithms and artificial intelligence (AI) tools are being used to automate decision-making and are having a significant impact on individuals’ rights and obligations. Controversies have emerged regarding the opaque nature of such schemes, the unintentional bias against and harm to underrepresented populations, and the broader legal, social, and ethical ramifications. State v. Loomis, a recent case in the United States, well demonstrates how unrestrained and unchecked outsourcing of public power to machines may undermine human rights and the rule of law. With a close examination of the case, this Article unpacks the issues of the ‘legal black box’ and the ‘technical black box’ to identify the risks posed by rampant ‘algorithmization’ of government functions to due process, equal protection, and transparency. We further assess some important governance proposals and suggest ways for improving the accountability of AI-facilitated decisions. As AI systems are commonly employed in consequential settings across jurisdictions, technologically-informed governance models are needed to locate optimal institutional designs that strike a balance between the benefits and costs of algorithmization.

Unione Europea 2019



Evoluzione della eXplainable AI - XAI

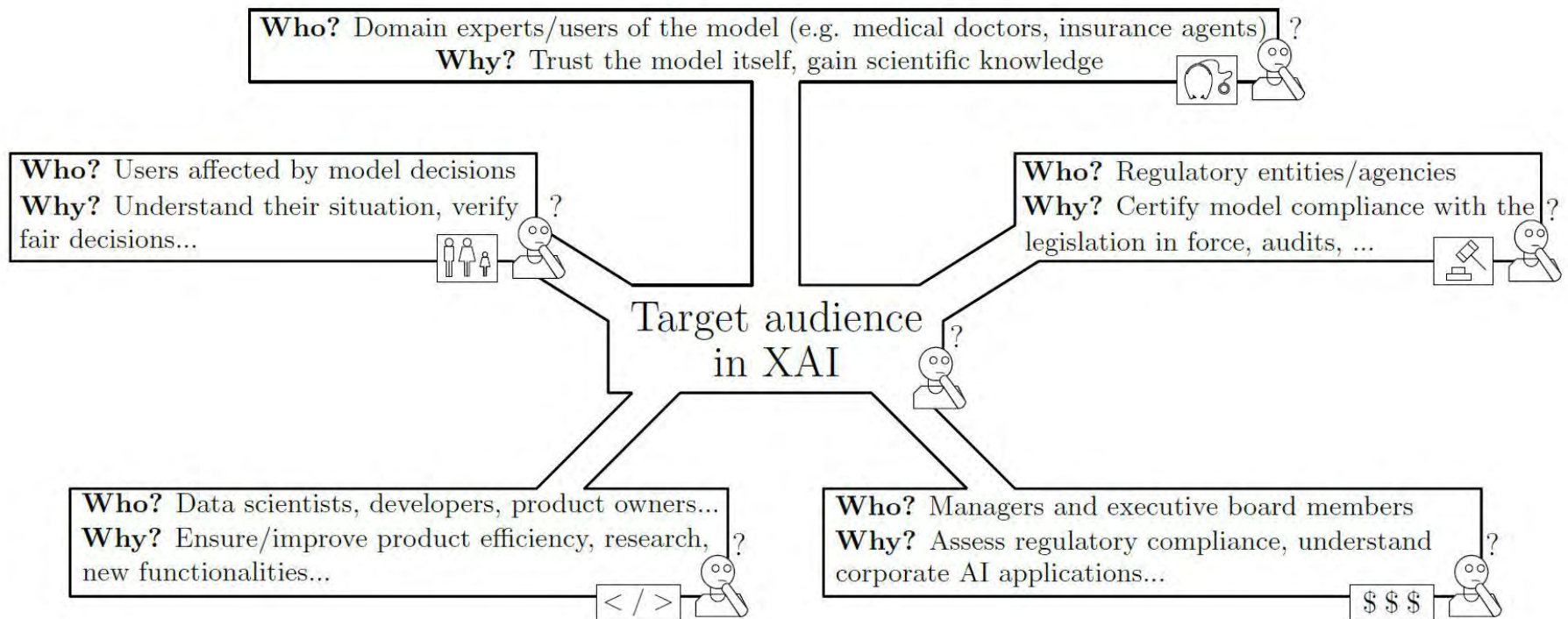
- **Interpretability**: it is defined as the ability to explain or to provide the meaning in understandable terms to a human.
- **Explainability**: explainability is associated with the notion of explanation as an interface between humans and a decision maker that is, at the same time, both an accurate proxy of the decision maker and comprehensible to humans



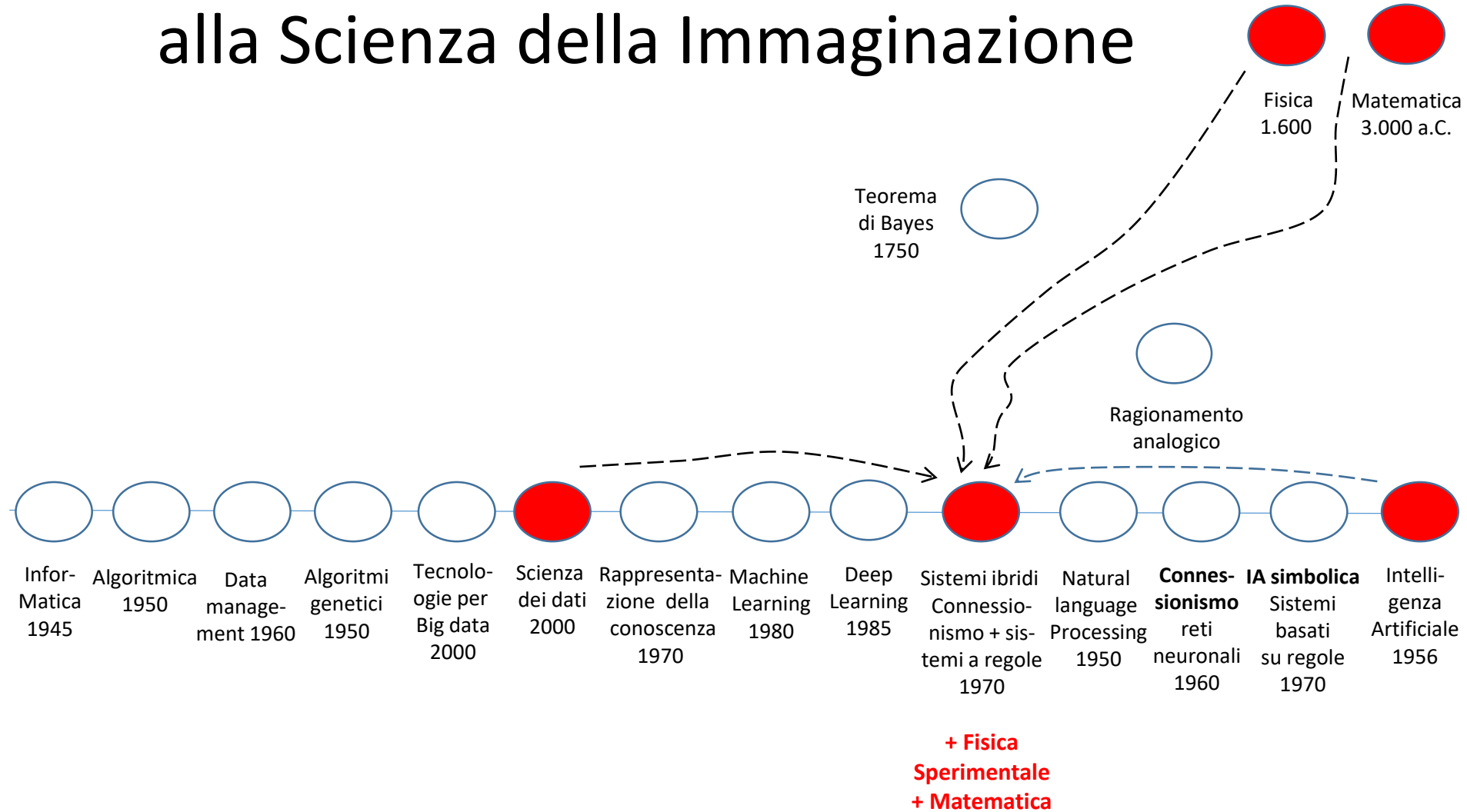
Termini usati nella comunità XAi

- **Understandability** (or equivalently, intelligibility) denotes the characteristic of a model to make human understand its function – how the model works – without any need for explaining its internal structure or the algorithmic means by which the model processes data internally
- **Comprehensibility**: when conceived for ML models, comprehensibility refers to the ability of a learning algorithm to represent its learned knowledge in a human understandable fashion. This notion of model comprehensibility stems from the postulates of Michalski, which stated that “the results of computer induction should be symbolic descriptions of given entities, semantically and structurally similar to those a human expert might produce observing the same entities. Components of these descriptions should be comprehensible as single ‘chunks’ of information, directly interpretable in natural language, and should relate quantitative and qualitative concepts in an integrated fashion”. Given its difficult quantification, comprehensibility is normally tied to the evaluation of the model complexity.
- **Interpretability**: it is defined as the ability to explain or to provide the meaning in understandable terms to a human.
- **Explainability**: explainability is associated with the notion of explanation as an interface between humans and a decision maker that is, at the same time, both an accurate proxy of the decision maker and comprehensible to humans.
- **Transparency**: a model is considered to be transparent if by itself it is understandable. Since a model can feature different degrees of understandability, transparent models in Section 3 are divided into three categories: simulatable models, decomposable models and algorithmically transparent models.

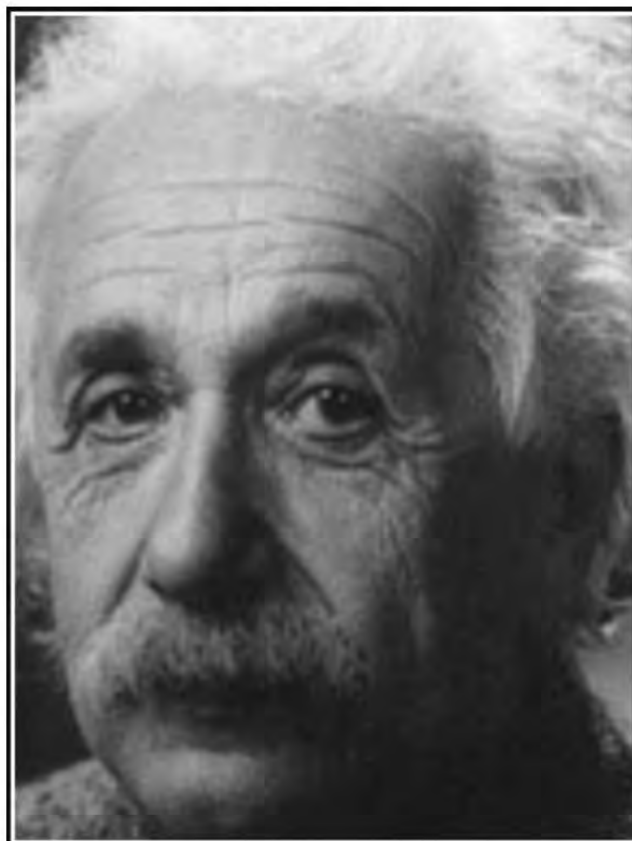
Le diverse forme e i diversi soggetti interessati alla Explainability



Dalla Scienza dei Dati alla Scienza della Immaginazione



Conoscere e immaginare



Imagination is more important than
knowledge. Knowledge is limited.
Imagination encircles the world.

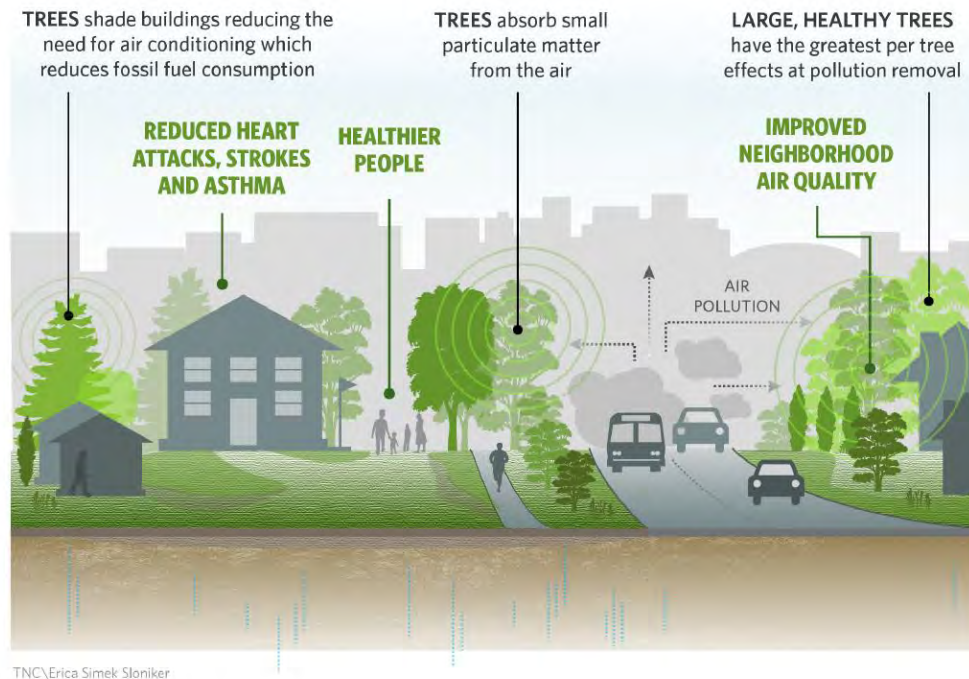
— *Albert Einstein* —

AZ QUOTES

Come intervenire per contrastare l'inquinamento? - 1

URBAN TREES, BETTER AIR QUALITY

Trees in cities can remove up to a quarter of the particulate matter pollution in their immediate vicinity. And when planted between a source of pollution and an apartment building, school or hospital, urban trees can help protect human health.



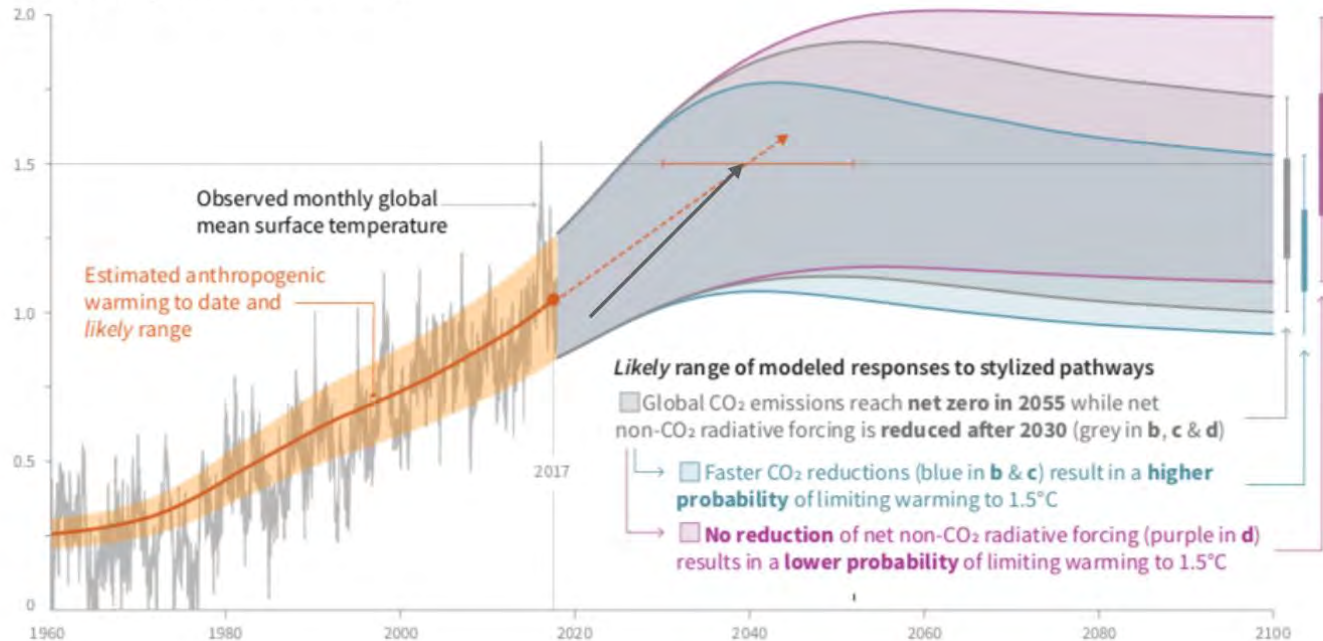
Come intervenire per contrastare l'inquinamento? - 2

UN Report on Climate Change October 2018

Cumulative emissions of CO₂ and future non-CO₂ radiative forcing determine the probability of limiting warming to 1.5°C

a) Observed global temperature change and modeled responses to stylized anthropogenic emission and forcing pathways

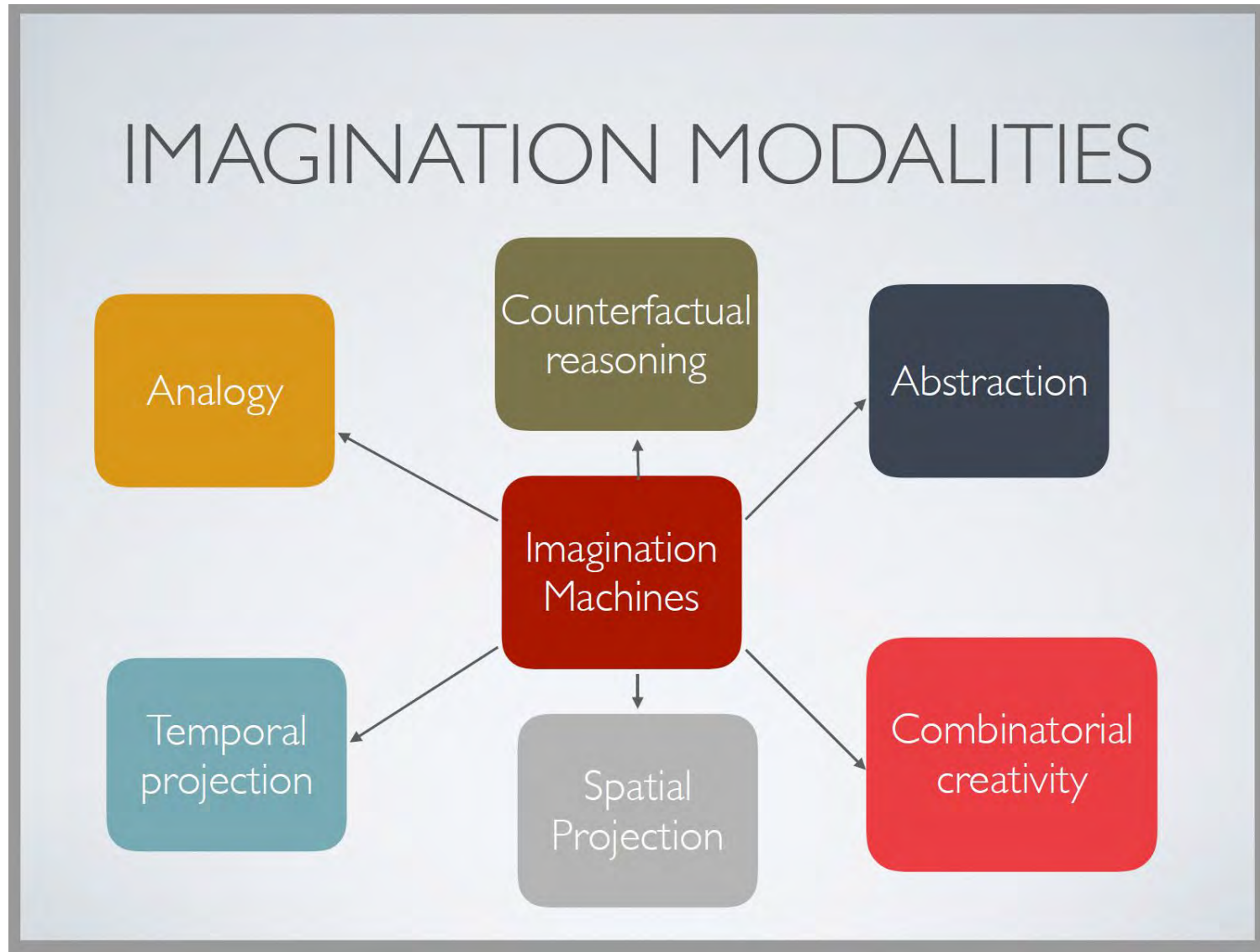
Global warming relative to 1850-1900 (°C)



Data Science

Imagination Science

I modi della immaginazione



Imagination Science

IMAGINATION SCIENCE: BEYOND DATA SCIENCE

Sridhar Mahadevan



UMass**Amherst**
College of Information
and Computer Sciences

AAAI Tutorial (January 28 1:30-5:30 p.m., 2019, Hawaii)

Le Adversarial networks

CAN: Creative Adversarial Networks Generating “Art” by Learning About Styles and Deviating from Style Norms*

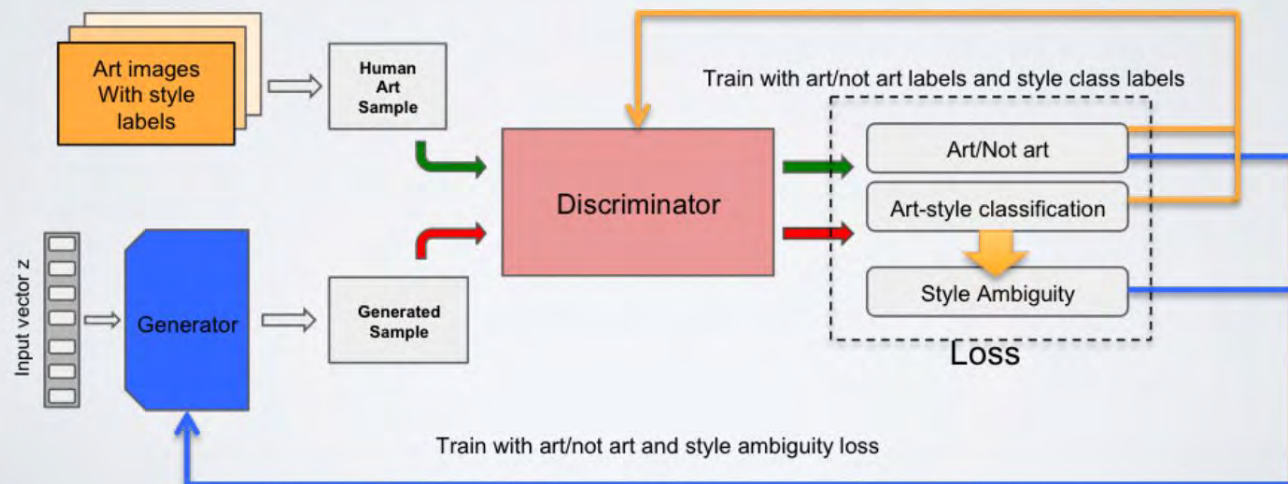
Ahmed Elgammal^{1†} Bingchen Liu¹ Mohamed Elhoseiny² Marian Mazzone³

The Art & AI Laboratory - Rutgers University

¹ Department of Computer Science, Rutgers University, NJ, USA

² Facebook AI Research, CA, USA

³ Department of Art History, College of Charleston, SC, USA



Il contributo della Topologia e della Fisica

FIBER BUNDLE IMAGINATION SPACE

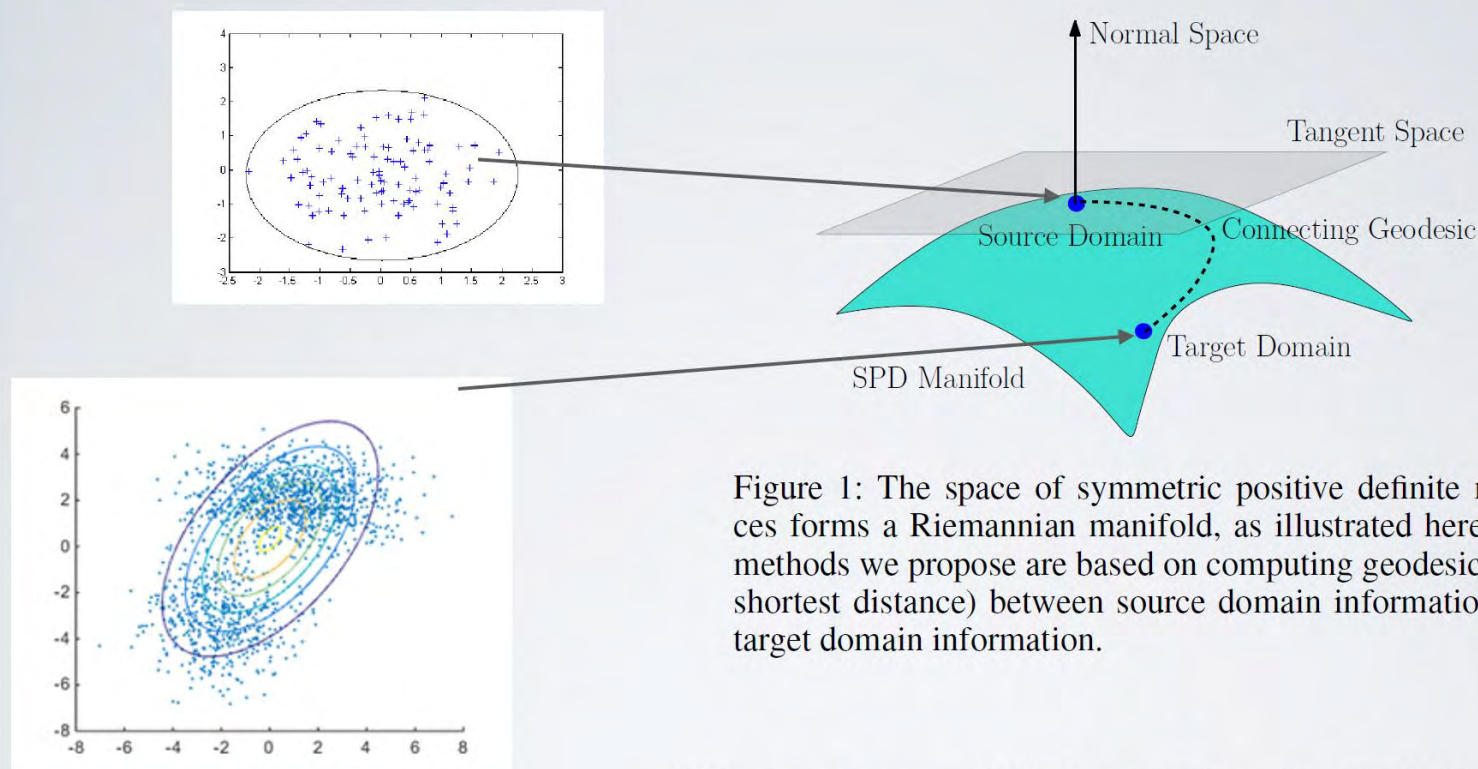


Figure 1: The space of symmetric positive definite matrices forms a Riemannian manifold, as illustrated here. The methods we propose are based on computing geodesics (the shortest distance) between source domain information and target domain information.

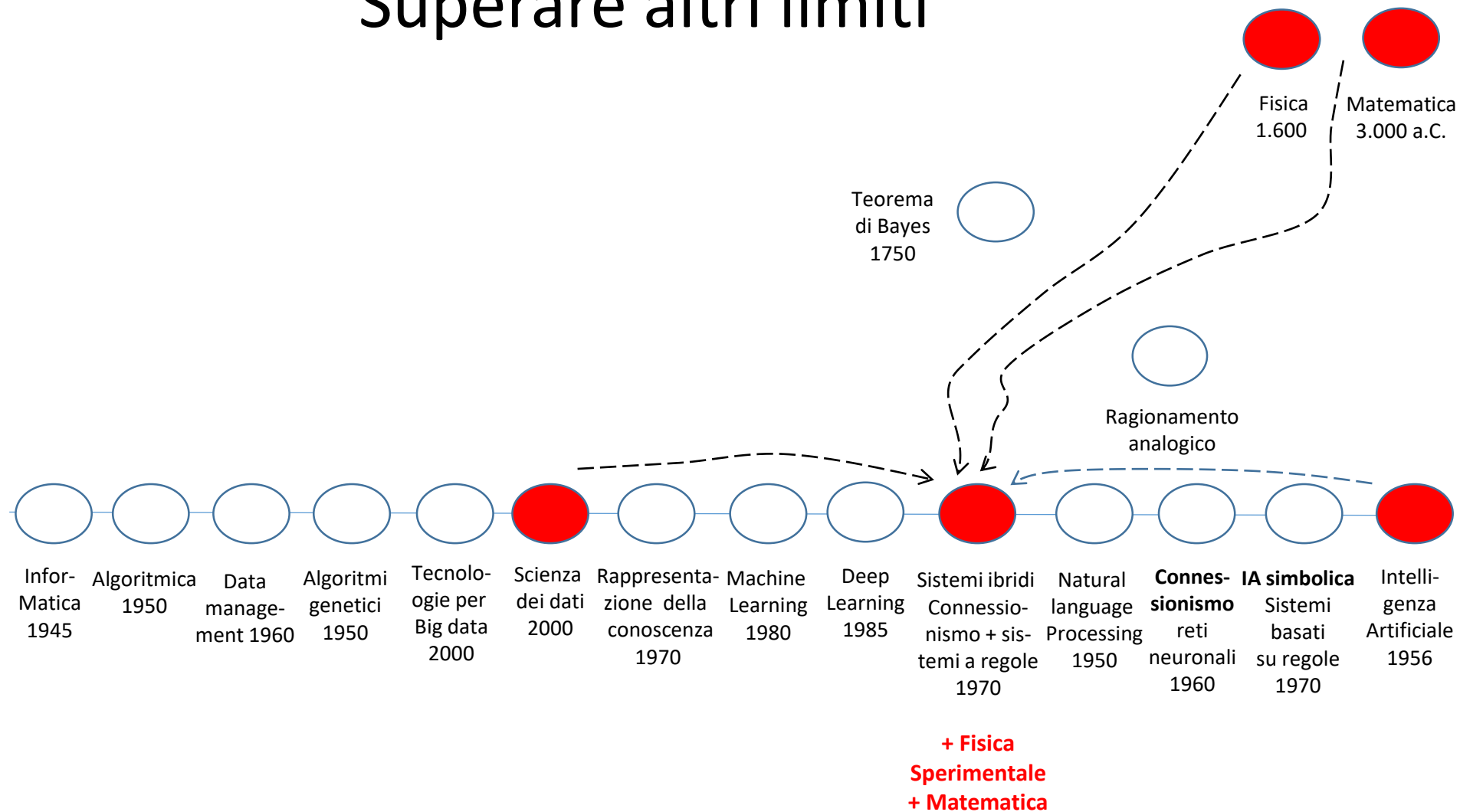
(Mahadevan, Mishra, Ghosh, ECML 2018)

Il contributo della Matematica

CAN loss function

$$\begin{aligned} \min_G \max_D V(D, G) = & \mathbb{E}_{x, \hat{c} \sim p_{data}} [\log D_r(x) + \log D_c(c = \hat{c}|x)] + \\ & \mathbb{E}_{z \sim p_z} [\log(1 - D_r(G(z))) - \sum_{k=1}^K \left(\frac{1}{K} \log(D_c(c_k|G(z))) + \right. \\ & \left. (1 - \frac{1}{K}) \log(1 - D_c(c_k|G(z))) \right)], \end{aligned}$$

Superare altri limiti



I limiti di Alexa

Amazon Alexa è un assistente personale intelligente sviluppato dall'azienda statunitense Amazon, utilizzato per la prima volta nei dispositivi Amazon



Ezra Hill
@ezrahill



Alexa, what time does IKEA open? IKEA opens at 9am..... I'm currently sat in the car park waiting until 10am [#echo](#) [#alexa](#) [#fail](#)

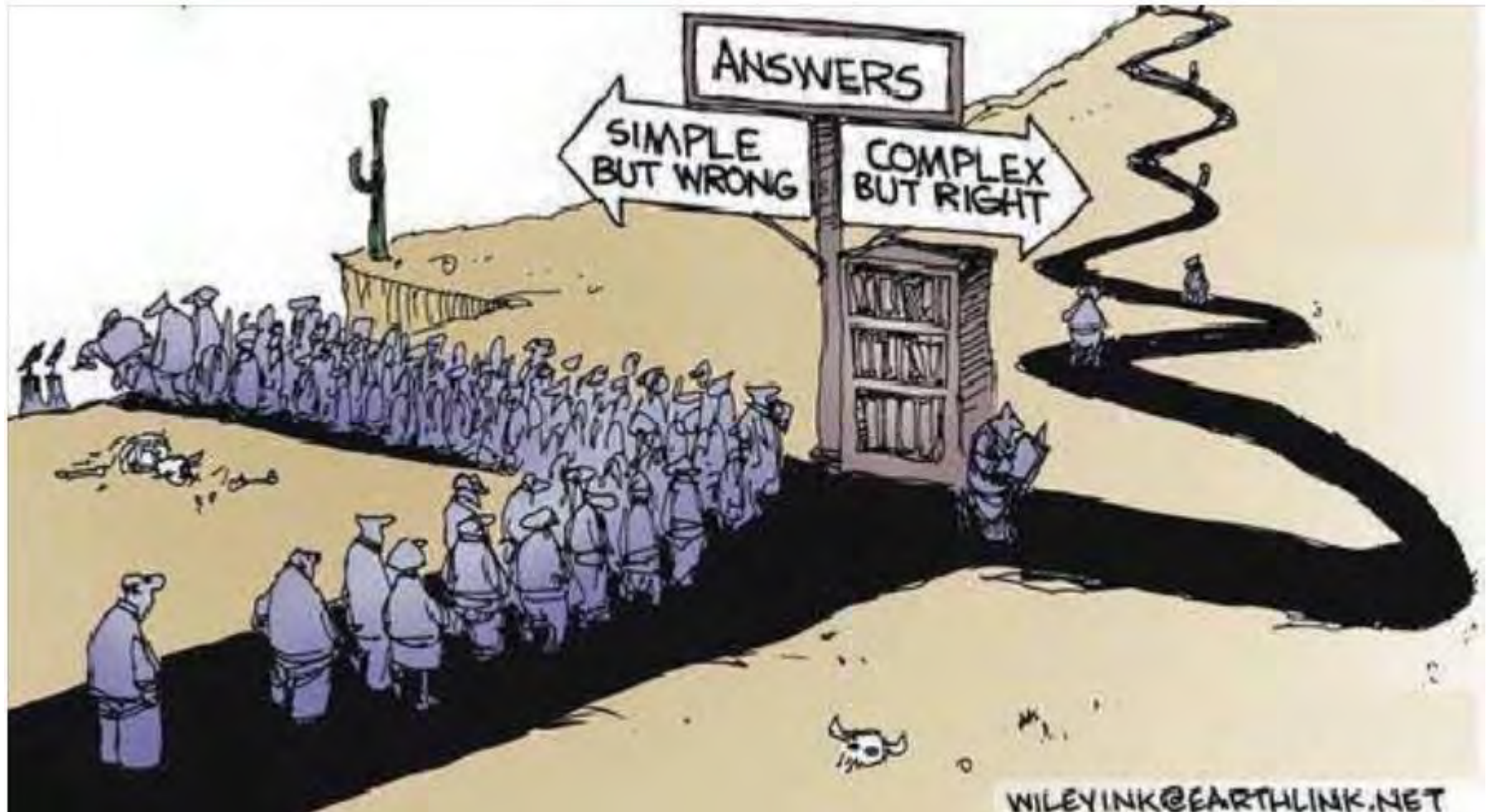
9:42 AM - Dec 2, 2017



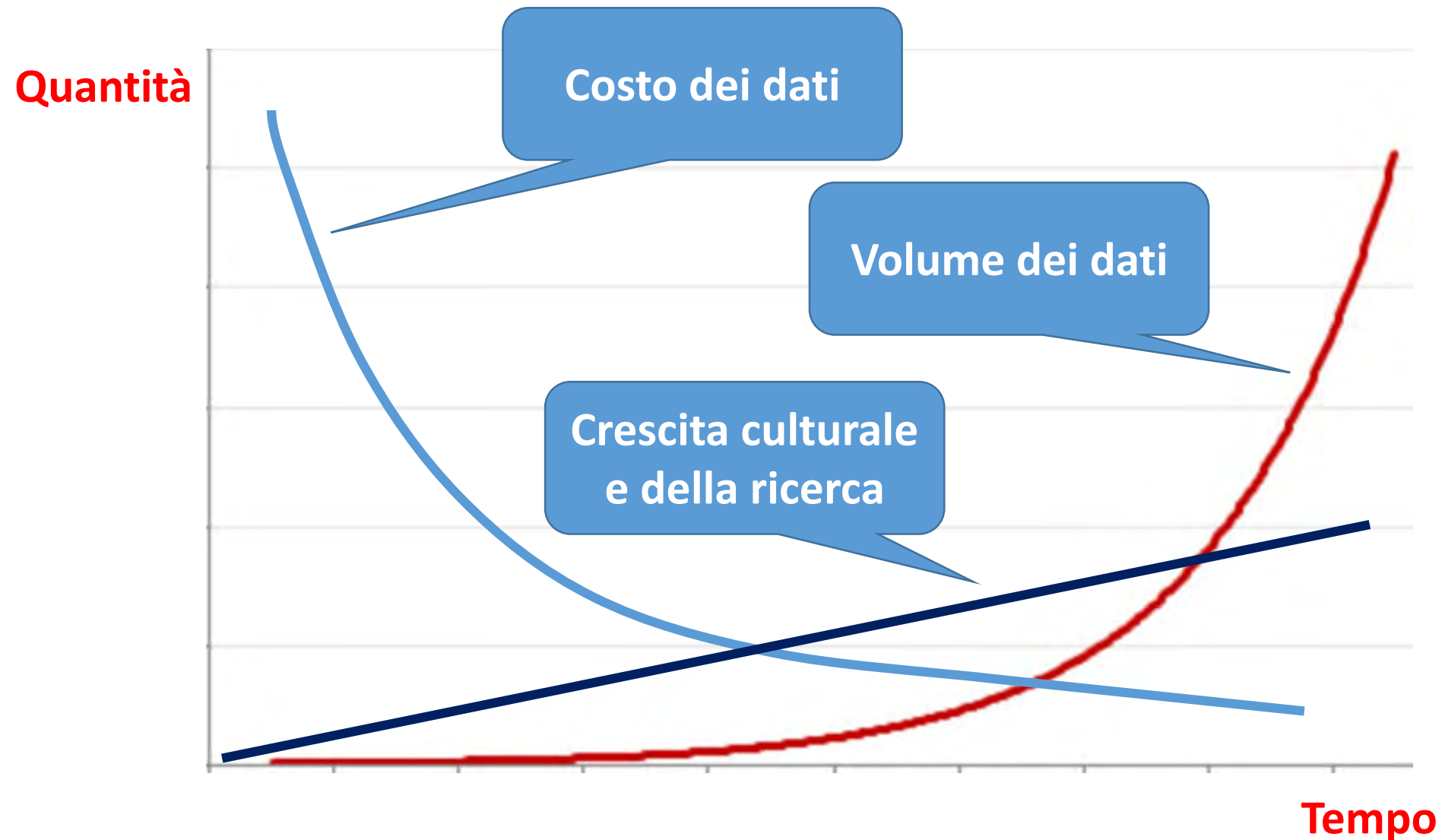
Le applicazione di IA richiedono più del ML :

- Modelli semantici
- Qualità dei dati
- Progettazione user centric

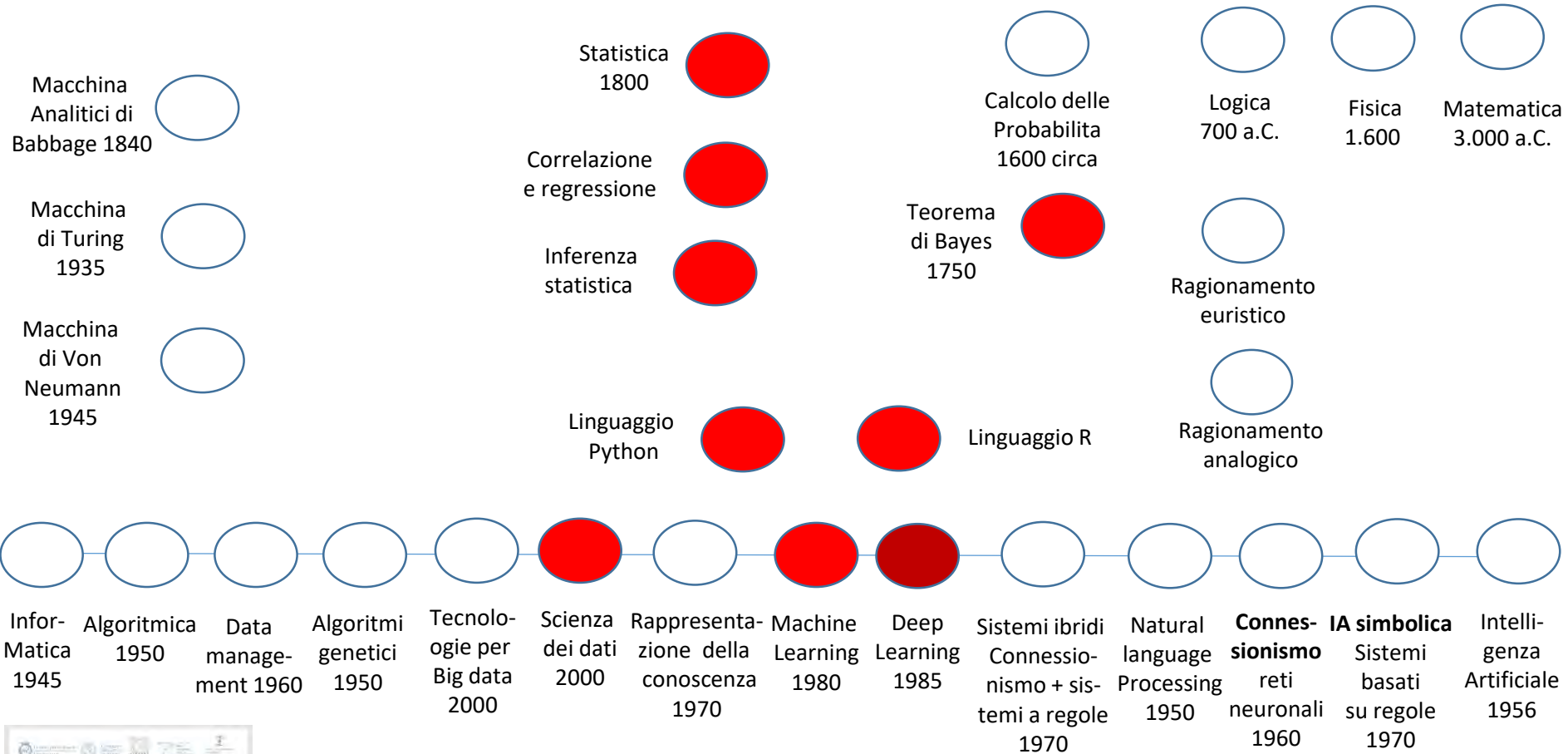
Non ci sono risposte semplici
a questioni complesse



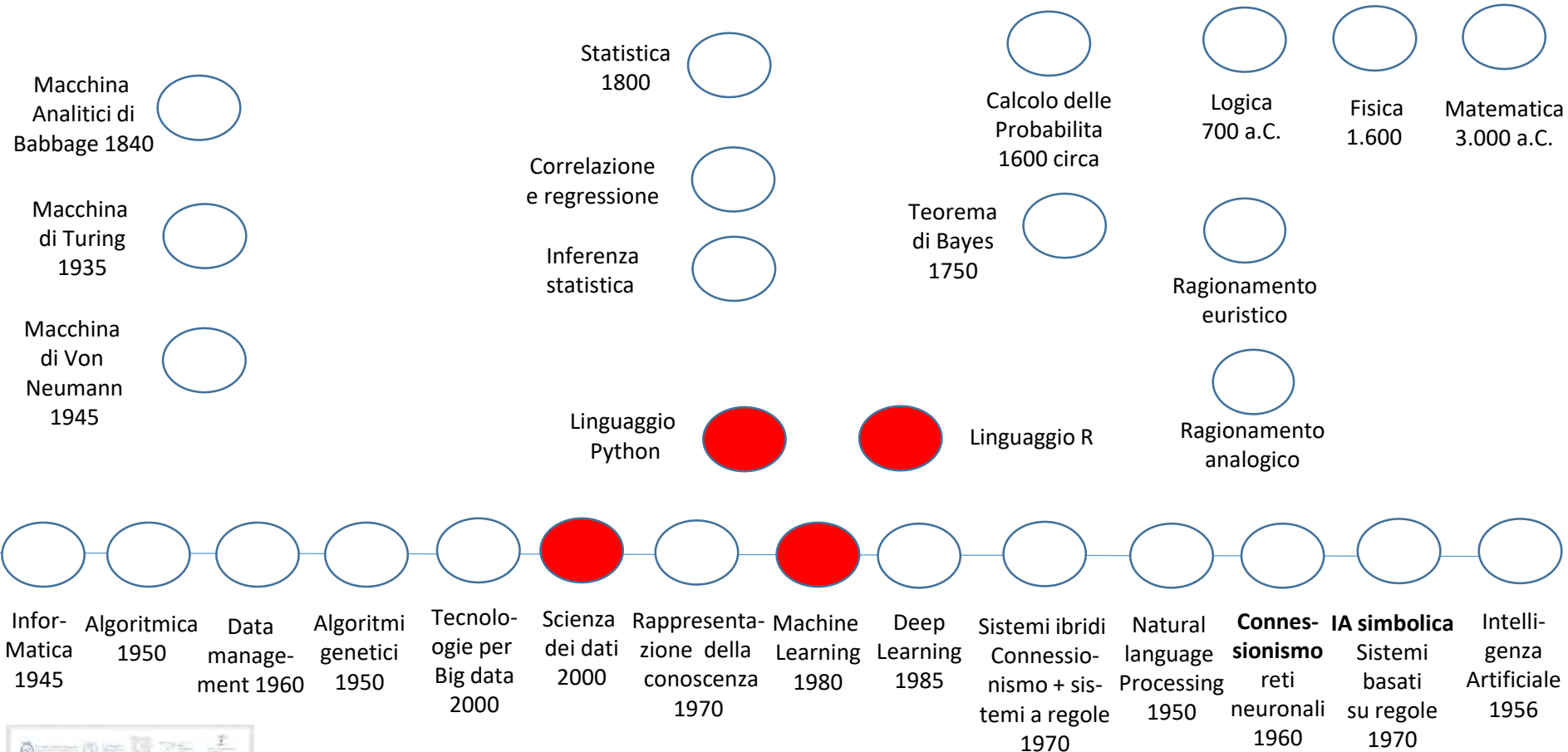
Diversa tendenza della crescita dei dati e della ricerca e cultura sui dati



Mappa dei saperi nel Corso



Corsi Federica Web Learning

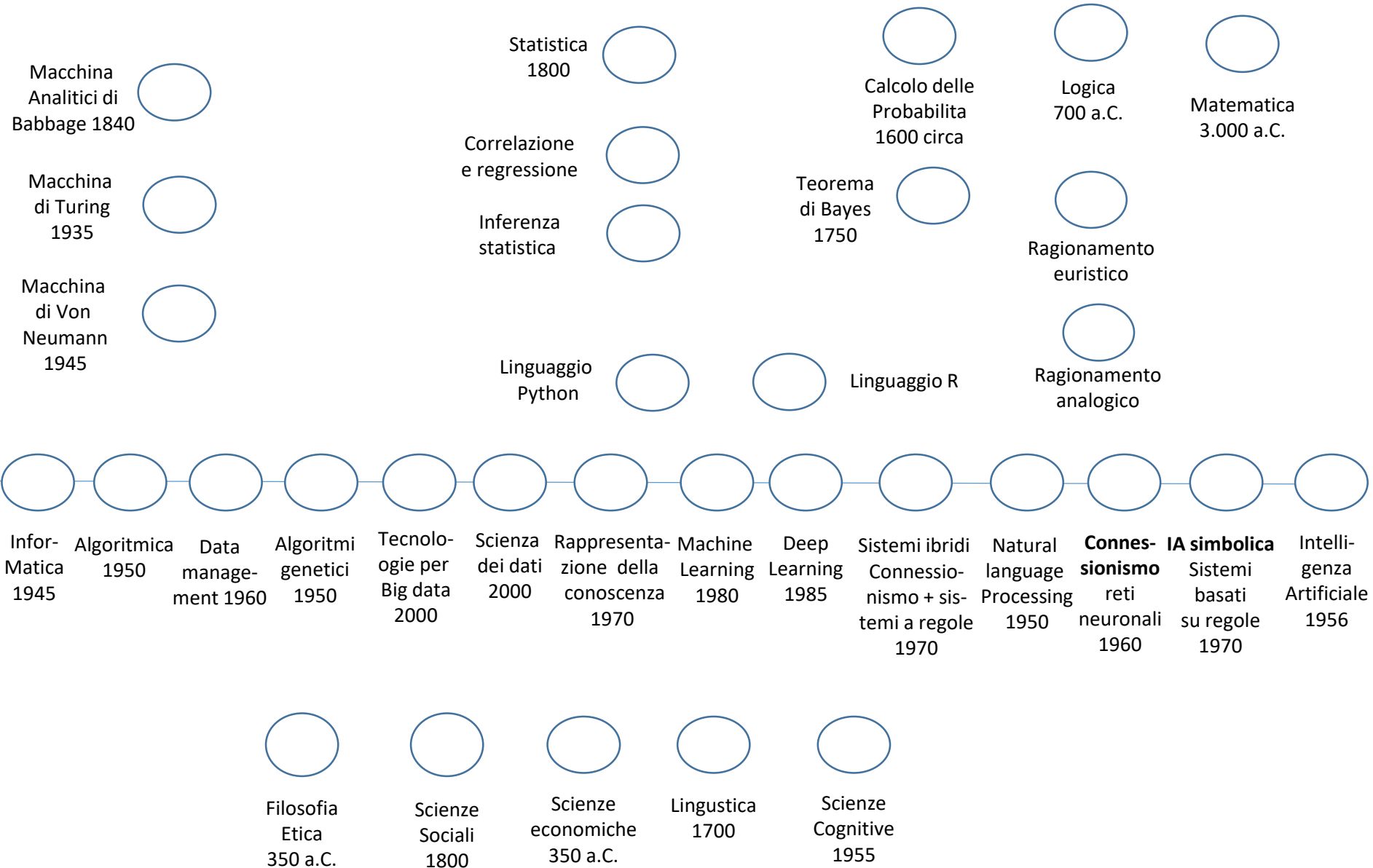


Riferimenti

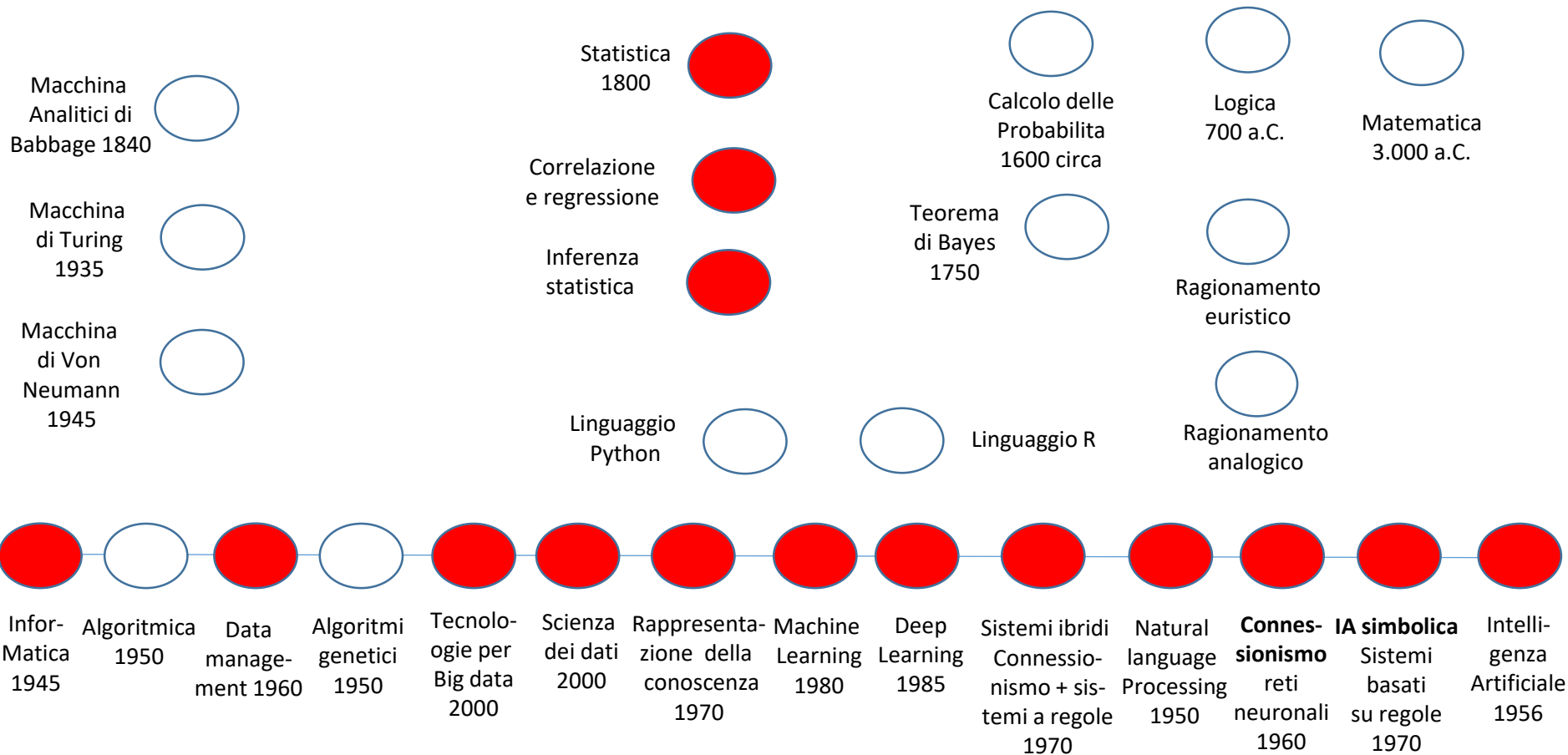
- P. Domingos – The Master Algorithm, Basic Books, 2015

Resti

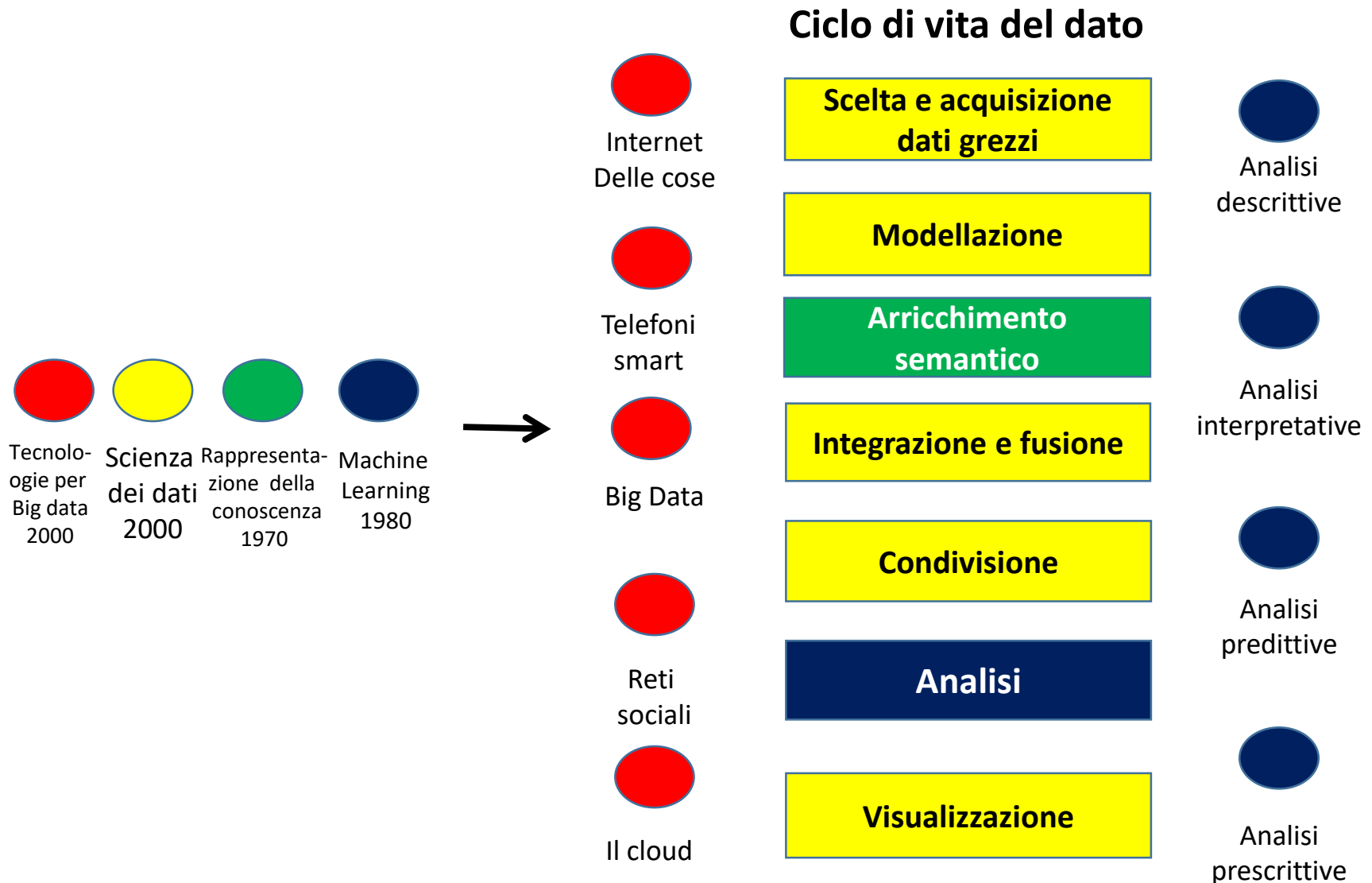
Mappa dei saperi nella Scienza dei dati



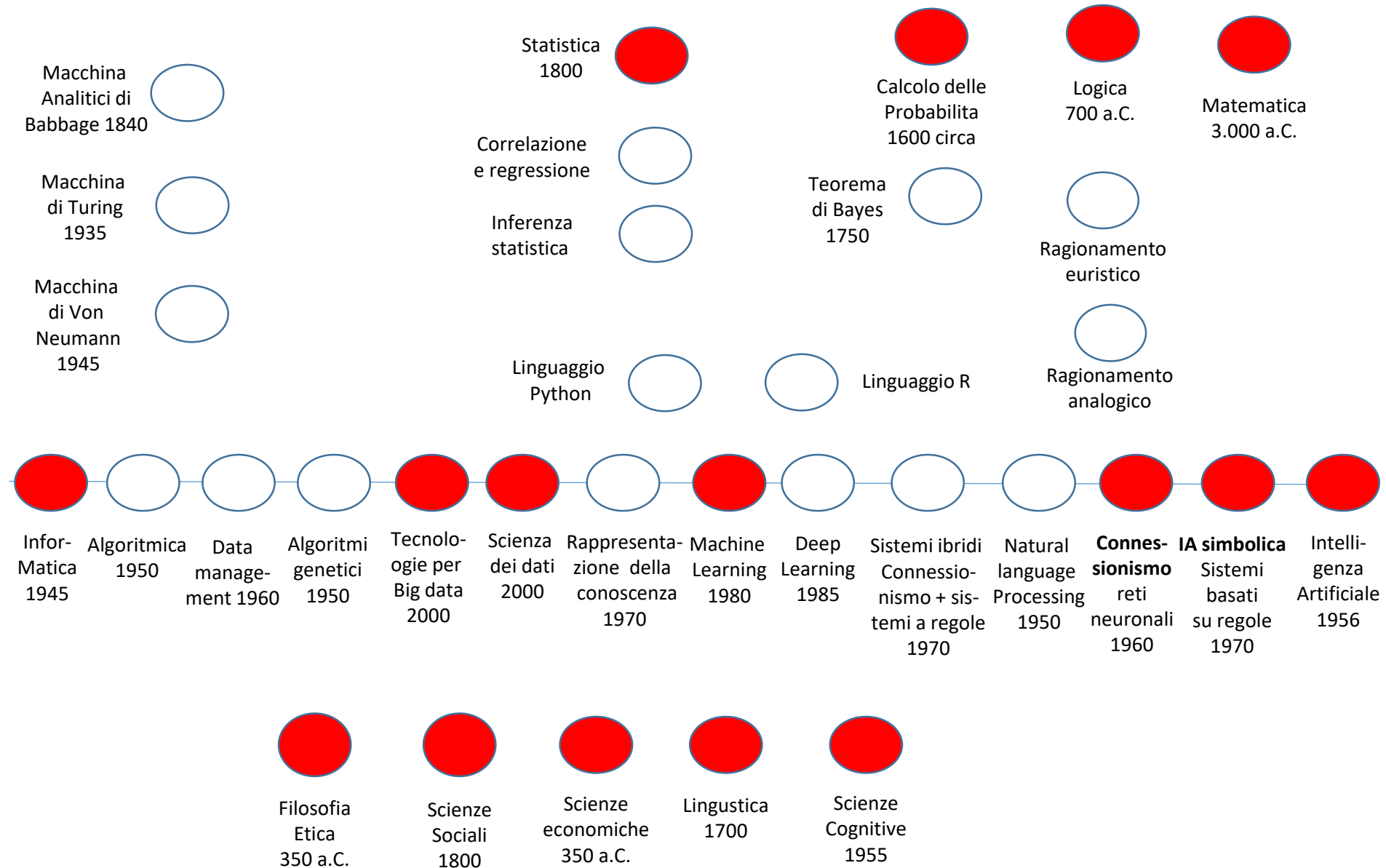
Scienza dei dati & Intelligenza Artificiale



La Scienza dei dati a volo d'uccello



Mappa dei saperi nella Scienza dei dati e nella Intelligenza Artificiale



Metodi basati su learning - 3

Ciò permette di associare l'output, nel nostro caso il livello di inquinamento, per ogni possibile input fornito dall'utente

latitudine	longitudine	Anno	Mese	Giorno	Ora	Minuto	Traffico	Riscaldamento	Qualità dell'aria	Livello di inquinamento
308.40	215.20	2017	6	12	8	10	0.7	0.7	47	3
308.40	215.20	2017	7	12	8	20	0.7	0.8	49	2
308.40	215.20	2017	6	12	8	30	0.8	0.8	51	5
308.40	215.20	2017	6	12	8	40	0.9	0.7	54	6
..	..	..	..	..	..	.			..	4
308.45	215.25	2019	12	12	8	10	0.7	0.7	54	4
308.45	215.25	2019	12	12	9	20	0.7	0.7	57	5
308.45	215.25	2019	12	12	10	30	0.8	0.8	60	6
308.45	215.25	2019	12	12	11	40	0.9	0.7	65	9
308.45	215.25	2019	12	12	12	50	0.8	0.6	70	6

Human level AI

1. Gli umani apprendono più velocemente delle machine spesso senza disporre di training data (che invece sono lo strumento di apprendimento tipico dell'apprendimento supervisionato, vedi in seguito)
2. Le differenti abilità cognitive degli umani sono altamente interconnesse e non sono come nelle macchine apprese e utilizzate indipendentemente
3. Gli umani possono immaginare e ragionare ad un elevato livello e profondità di astrazione e su cose ed eventi che non esistono

Google Translator



translator



Tutti

Notizie

Immagini

Video

Maps

Altro

Impostazioni

Strumenti

Circa 234.000.000 risultati (0,35 secondi)

Italiano ▼



Inglese ▼

Garibaldi nacque
a Nizza



Garibaldi was born in
Nice



Apri in Google Traduttore

Feedback