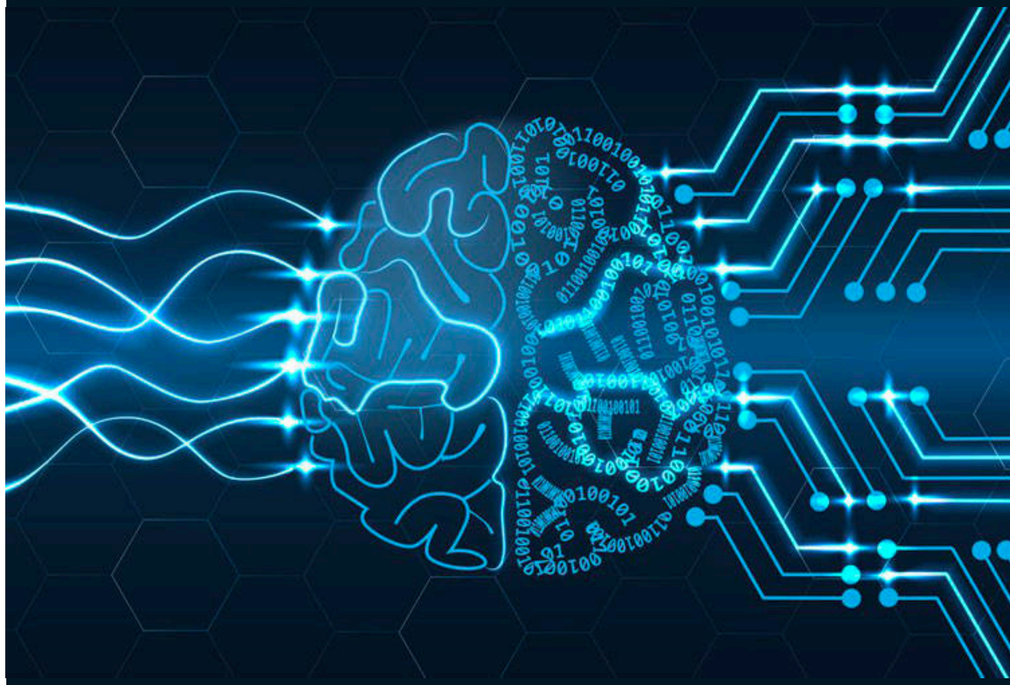
FONDAZIONE BRUNO KESSLER

MPBA PREDICTIVE MODELS FOR BIOMEDICINE & ENVIRONMENT

TIFPA Trento Institute for Fundamental Physics and Applications

Machine Learning and Data Science

M. Cristoforetti
Fondazione Bruno Kessler - Trento, TIFPA

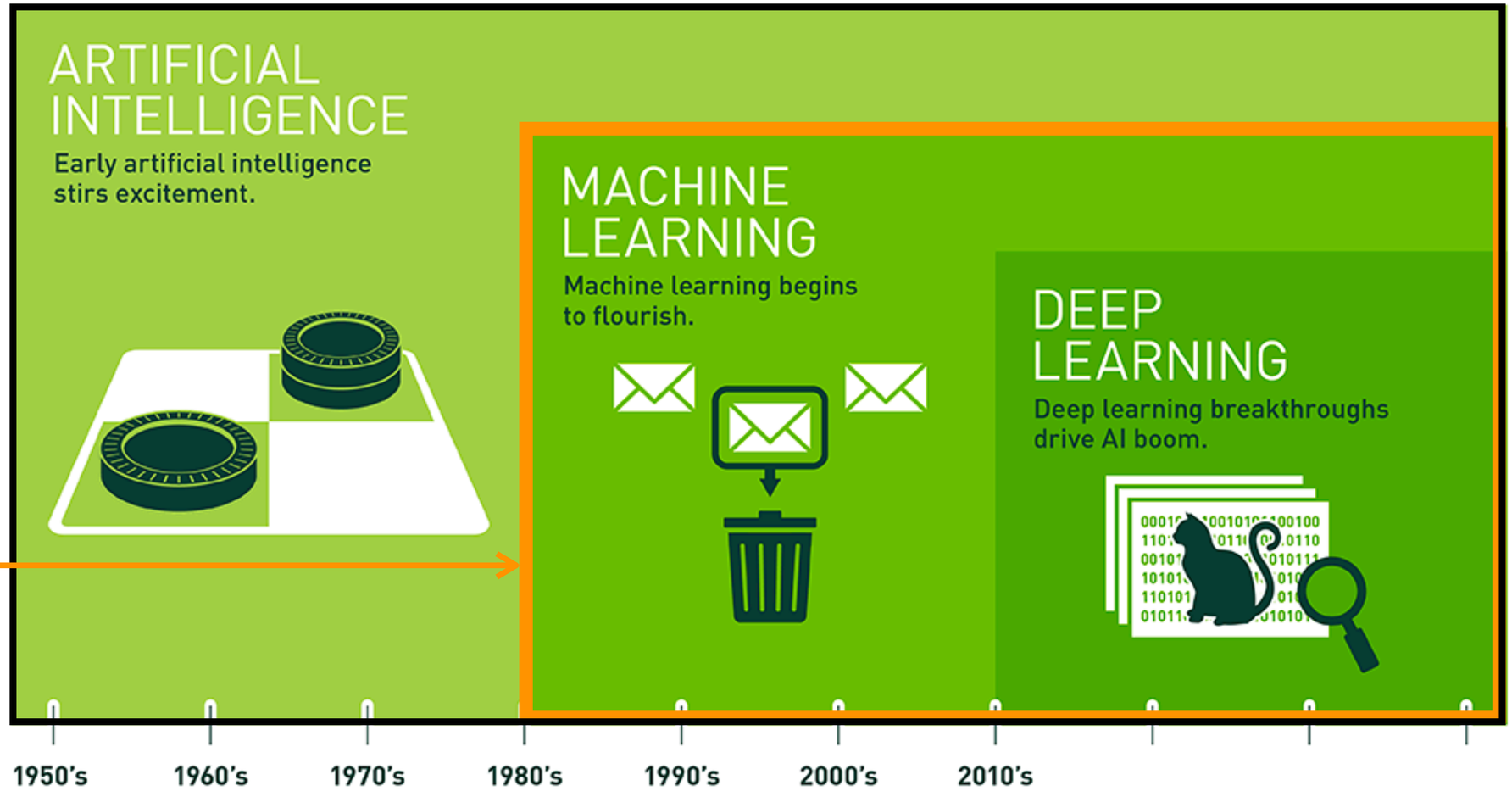


Gran Sasso Science Institute - 25 Ottobre 2018

DI COSA PARLIAMO ?

Focus:

DATA SCIENCE



FONDAZIONE BRUNO KESSLER

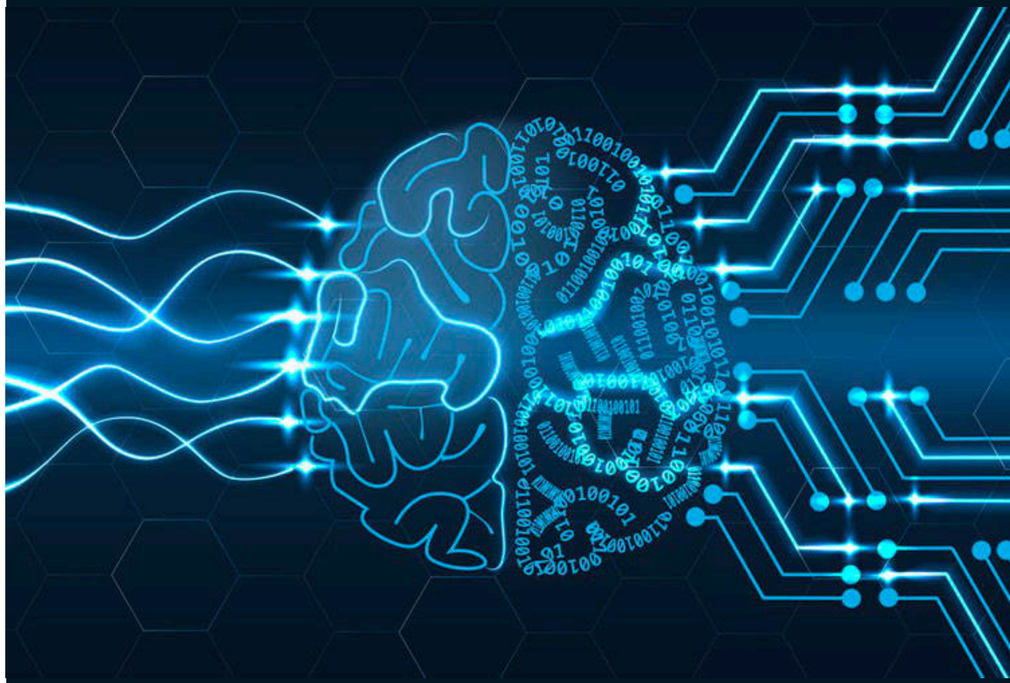
MPBA PREDICTIVE MODELS FOR BIOMEDICINE & ENVIRONMENT

TIFPA Trento Institute for Fundamental Physics and Applications

Machine Learning

~~and~~ for

Data Science



IL METODO SCIENTIFICO

1. Osservazioni sistematiche
2. Formulazione di una domanda
3. Elaborazione di una ipotesi (teoria) soluzione della domanda
4. Validazione delle previsioni (e quindi della teoria) tramite nuovi esperimenti ed osservazioni

G. Bezzuoli, 1841

OSSERVAZIONI SISTEMATICHE

Che cosa osserva un data scientist?

MULTIDISCIPLINARIETA'

OSSERVAZIONI SISTEMATICHE

Grazie alla digitalizzazione e alle nuove tecnologie

la disponibilità di **DATI** aumenta in maniera esponenziale

DATI facilmente **accessibili** da governi, pubblica amministrazione, imprese

ma **anche** da **NOI**

1.1 MILIARDI di viaggi in taxi a NYC (2009-2015)

FONDAZIONE BRUNO KESSLER

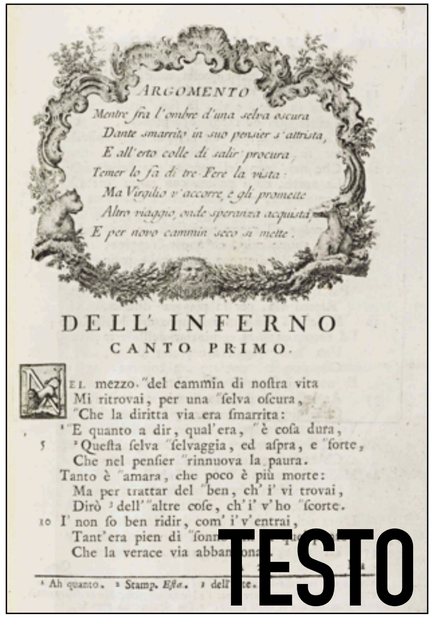
MPBA PREDICTIVE MODELS FOR BIOMEDICINE & ENVIRONMENT

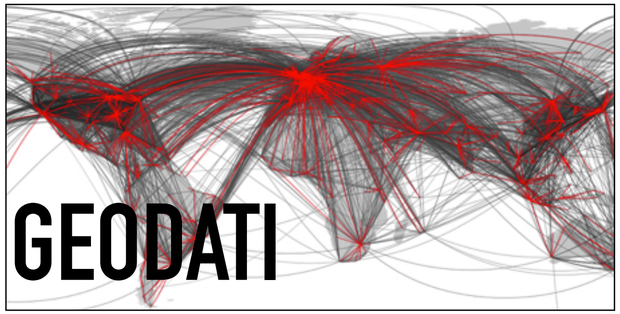
TIFPA Trento Institute for Fundamental Physics and Applications

OSSERVAZIONI SISTEMATICHE

DATI

ETEROGENEI in gran parte NON STRUTTURATI

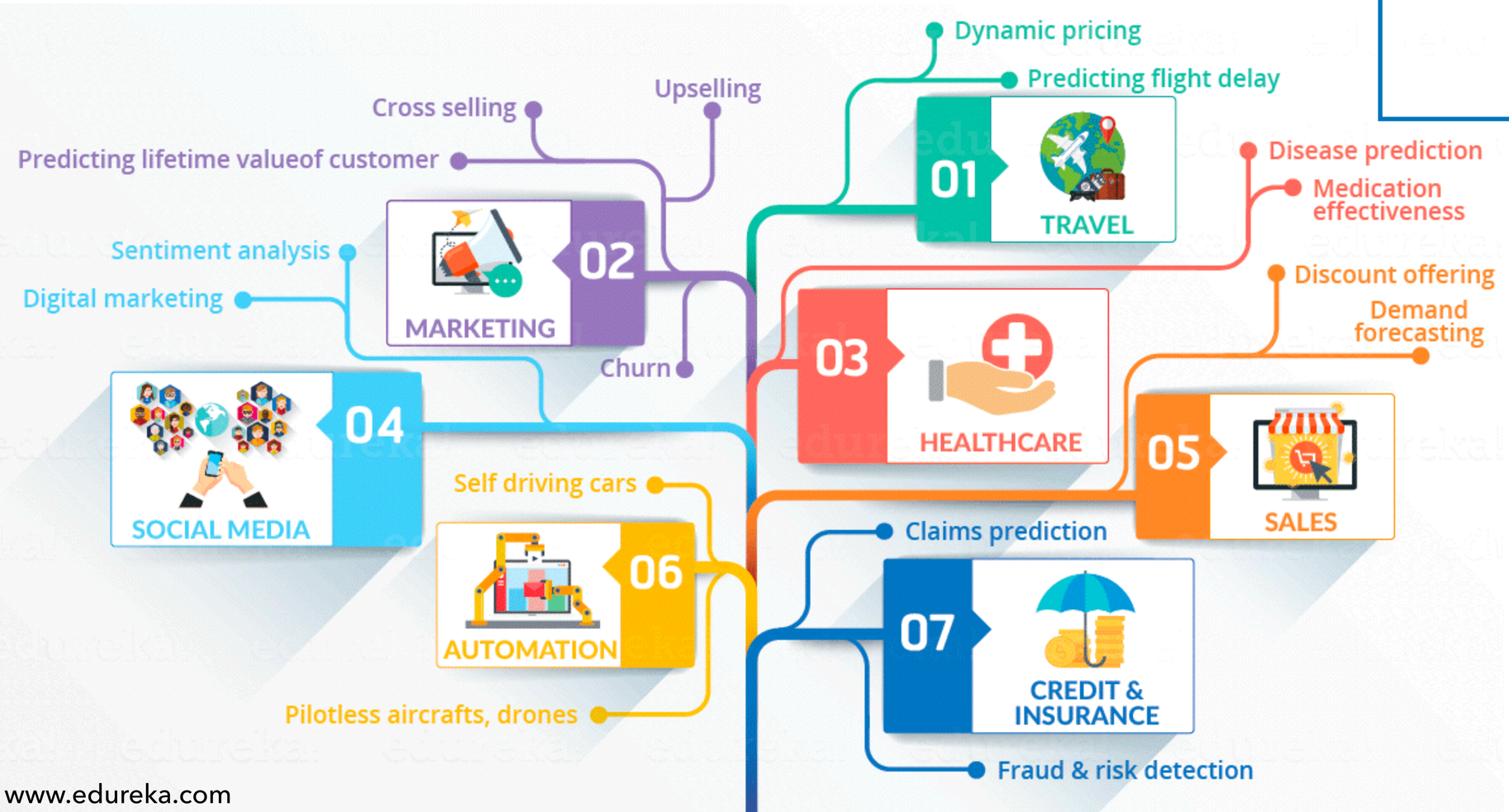




	A	B	C	D	E
1	Last Name	Sales	Country	Quarter	
2	Smith	$16,753.00	UK	Qtr 3	
3	Johnson	$14,808.00	USA	Qtr 4	
4	Williams	$10,644.00	UK	Qtr 2	
5	Jones	$1,390.00	USA	Qtr 3	
6	Brown	$4,865.00	USA	Qtr 4	
7	Williams	$12,438.00	UK	Qtr 1	

FORMULAZIONE DI UNA DOMANDA

(digital) **DATA DRIVEN SCIENCE**



Non basta avere **I DATI**

Si deve arrivare ad una **DOMANDA**

ELABORAZIONE DELLA TEORIA E VALIDAZIONE

Gli ingranaggi (**ALGORITMI**) sono gli stessi indipendentemente dalla disciplina.

Cambia il modo di assemblarli (**MODELLO**)

E ovviamente i dati a cui il modello è applicato

MULTIDISCIPLINARIETA'

Focalizzazione sulla **PREDITTIVITA'**

posponendo comprensione del fenomeno in termini di principi primi ("mattoni" fondamentali della disciplina)*

*questo non significa che poi io non mi interroghi sul perché il mio modello funzioni e quindi non arrivi alla comprensione

ELABORAZIONE DELLA TEORIA E VALIDAZIONE

Approccio Data Science

Maggiore predittività

Interpretazione del modello

Nature 560, 632–634 (2018)

LETTER

https://doi.org/10.1038/s41586-018-0438-y

Deep learning of aftershock patterns following large earthquakes

Phoebe M. R. DeVries[1,2*], Fernanda Viégas[3], Martin Wattenberg[3] & Brendan J. Meade[1]

Aftershocks are a response to changes in stress generated by large earthquakes and represent the most common observations of the triggering of earthquakes. The maximum magnitude of aftershocks and their temporal decay are well described by empirical laws (such as Bath's law[1] and Omori's law[2]), but explaining and forecasting the spatial distribution of aftershocks is more difficult. Coulomb failure stress change[3] is perhaps the most widely used criterion to explain the spatial distributions of aftershocks[4–8], but its applicability has been disputed[9–11]. Here we use a deep-learning approach to identify a static-stress-based criterion that forecasts aftershock locations without prior assumptions about fault orientation. We show that a neural network trained on more than 131,000 mainshock–aftershock pairs can predict the locations of aftershocks in an independent test dataset of more than 30,000 mainshock–aftershock pairs more accurately (area under curve of 0.849) than can classic Coulomb failure stress change (area under curve of 0.583). We find that the learned aftershock pattern is physically interpretable: the maximum change in shear stress, the von Mises yield criterion (a scaled version

neuron may be interpreted as the predicted probability that a grid cell generates one or more aftershocks.

The stress changes and aftershock locations associated with about 75% of randomly selected distinct mainshocks were used as training data; the remaining 25% were reserved to test the trained neural networks. The training and testing datasets both consist of the elements of the stress-change tensor as features and the corresponding labels of either 0, for grid cells without aftershocks, or 1, for grid cells with aftershocks.

We assess the accuracy of the neural-network aftershock location forecasts on the test dataset using receiver operating characteristic (ROC) analysis. ROC curves are widely used to assess the efficacy of diagnostic medical tests. To build these curves, the true positive rate of a binary classifier is plotted against the false positive rate for all possible thresholds of the classifier (see Methods for more details). The area under an ROC curve (AUC) then quantifies the overall performance of a test across all thresholds (Fig. 1). The ROC analysis reveals that the neural-network forecast can explain aftershock locations better than can

ELABORAZIONE DELLA TEORIA E VALIDAZIONE

Gli ingranaggi (**ALGORITMI**) sono gli stessi indipendentemente dalla disciplina.

Cambia il modo di assemblarli (**MODELLO**)

MULTIDISCIPLINARIETA'

MACHINE LEARNING

MACHINE LEARNING

sviluppo di algoritmi e tecniche che permettano ai computer di **imparare**, deducendo dai dati osservati un modello che possa essere usato per la predizione.

PERCHE' ?

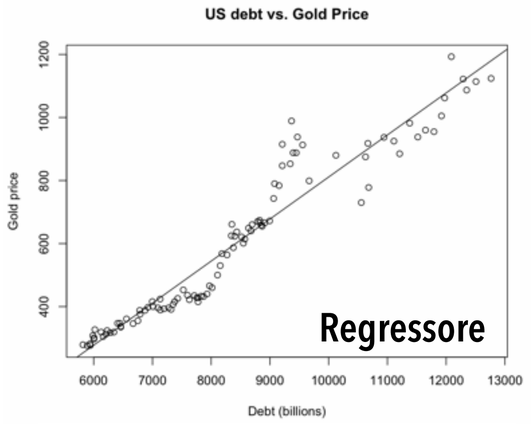
COMPLESSITA' del problema:
non siamo in grado di partire da leggi generali per dedurre il caso particolare

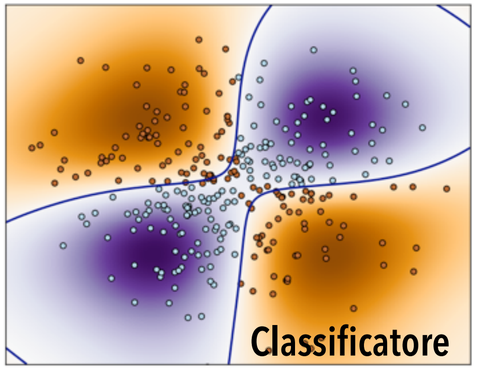
FLESSIBILITA' dell'algoritmo:
possiamo affrontare problemi in diverse discipline con lo stesso strumento

TIPI DI APPRENDIMENTO

Supervised Learning

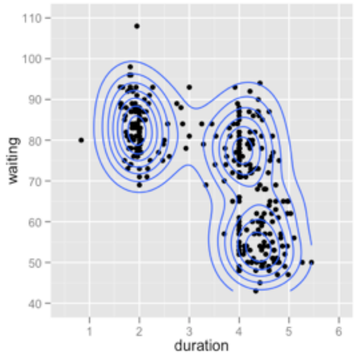
I dati hanno una "etichetta"





Unsupervised Learning

I dati NON hanno una "etichetta"

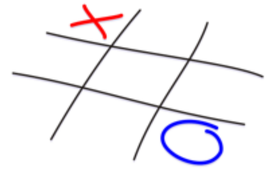


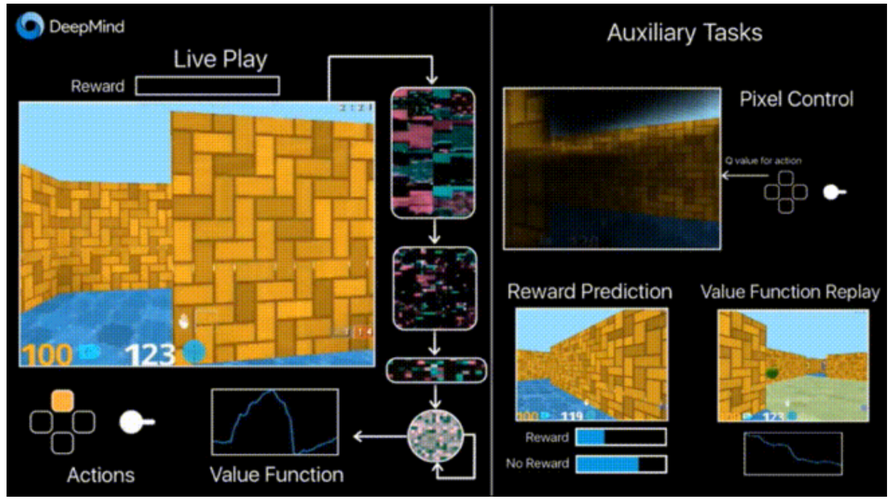
Stima densità

Clustering: class discovering

Reinforcement Learning

La macchina agisce per massimizzare un guadagno





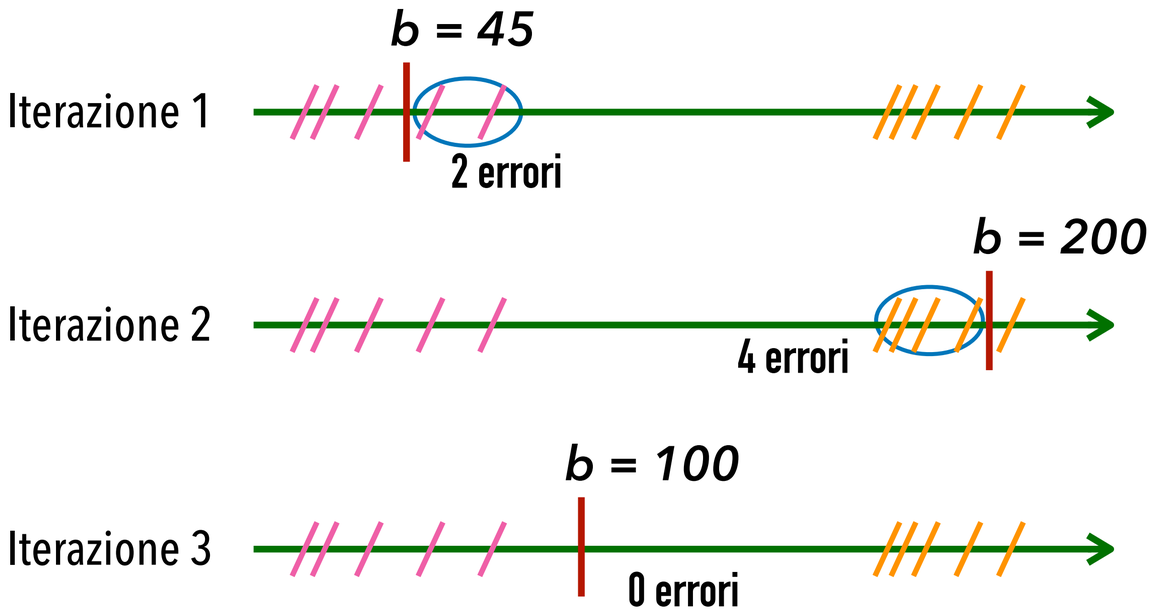
UNREAL (Unsupervised Reinforcement and Auxiliary Learning)

COME FUNZIONA ? Classificatore di girasoli

Altezza	Fiore	Label
180	Girasole	1
12	Altro	0
205	Girasole	1
23	Altro	0
50	Altro	0
77	Altro	0
178	Girasole	1
195	Girasole	1
36	Altro	0
183	Girasole	1

↑ Input ↑ Output

Modello a un parametro ***b***
L' algoritmo deve fissare ***b*** in modo che
data una nuova altezza (nuovo dato di input)
sappia dire se la pianta associata è un
girasole (label =1) oppure altro (label = 0)

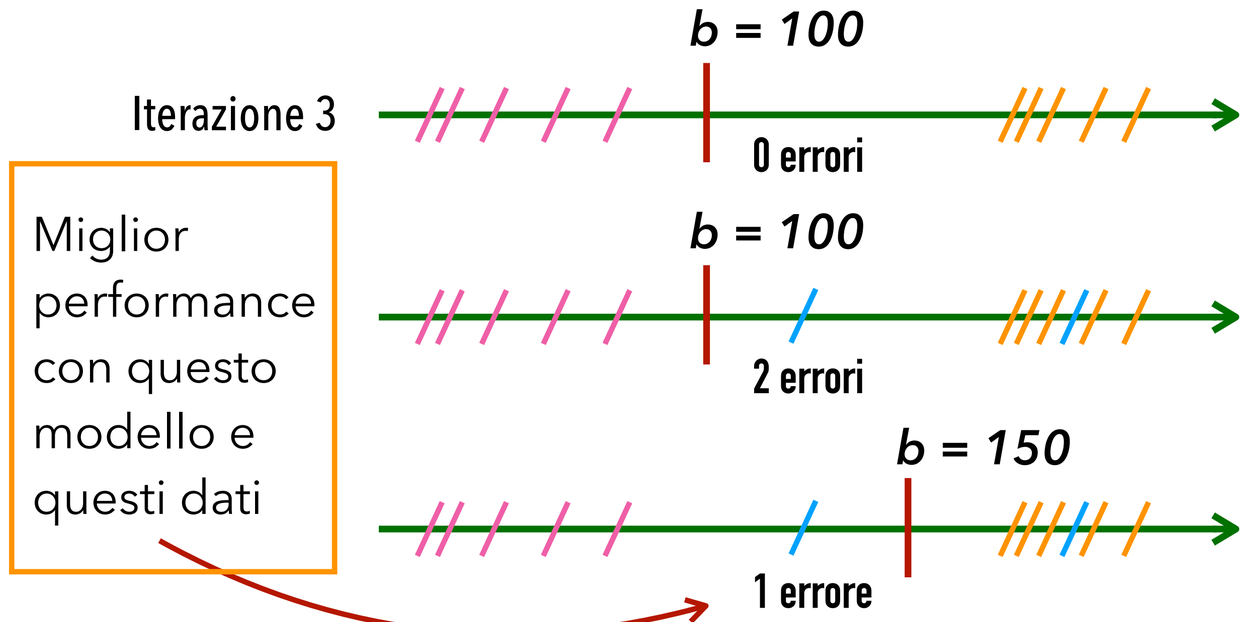


COME FUNZIONA ? Classificatore di girasoli

Altezza	Fiore	Label
180	Girasole	1
12	Altro	0
205	Girasole	1
23	Altro	0
50	Altro	0
77	Altro	0
178	Girasole	1
195	Girasole	1
36	Altro	0
183	Girasole	1
190	**Nuovo**	**0**
130	**Nuovo**	**0**

↑ Input ↑ Output

Modello a un parametro ***b***
L' algoritmo deve fissare ***b*** in modo che
data una nuova altezza (nuovo dato di input)
sappia dire se la pianta associata è un
girasole (label =1) oppure altro (label = 0)



COME FUNZIONA ?

L'algoritmo non sa nulla di piante, può essere usato ogni volta che voglio separare due classi utilizzando una variable che ritengo discriminante

Classificatore di girasoli

Modello a un parametro ***b***
L' algoritmo deve fissare ***b*** in modo che
data una nuova altezza (nuovo dato di input)
sappia dire se la pianta associata è un
girasole (label =1) oppure altro (label = 0)

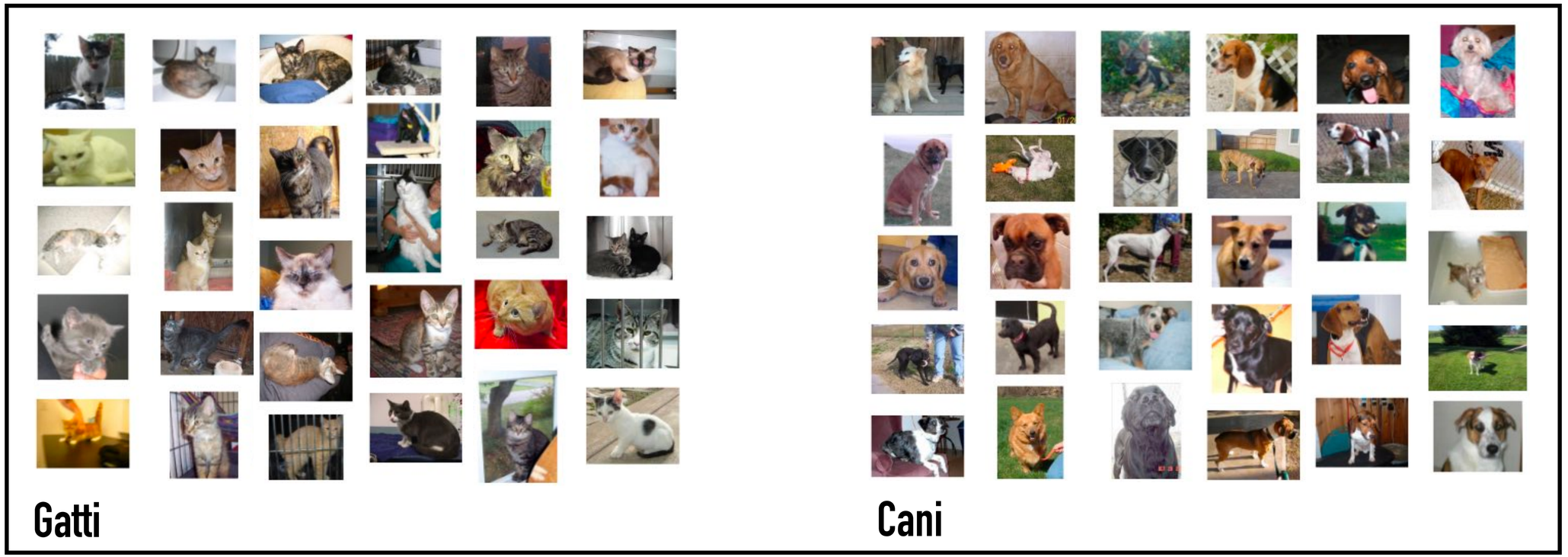
Classificatore di macchine sportive

Modello a un parametro ***b***
L' algoritmo deve fissare ***b*** in modo che
data una nuova velocità massima (nuovo dato di input)
sappia dire se la macchina associata è una
macchina sportiva (label =1) oppure altro (label = 0)

ELABORAZIONE DELLA TEORIA E VALIDAZIONE

Classificatore di **cani** e **gatti**

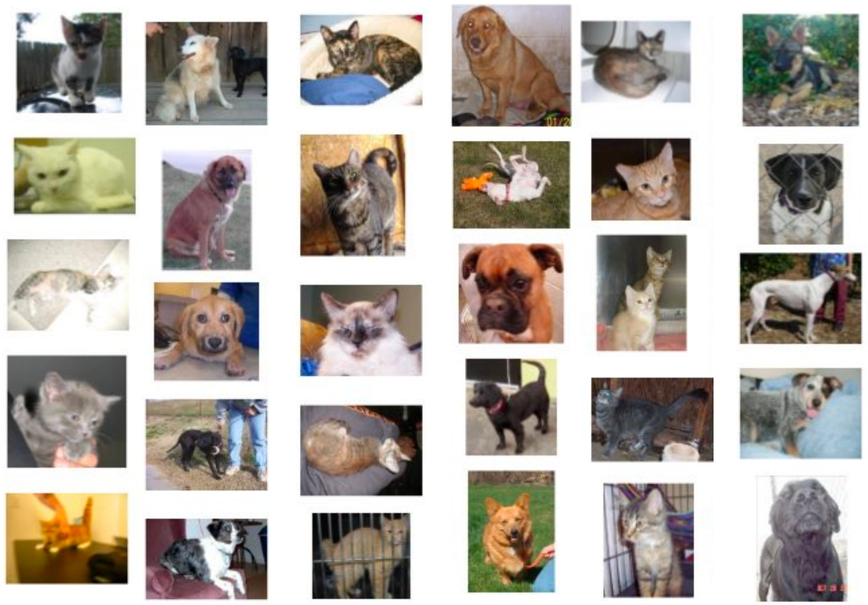
DATASET



ELABORAZIONE DELLA TEORIA E VALIDAZIONE

Classificatore di **cani** e **gatti**

TRAIN

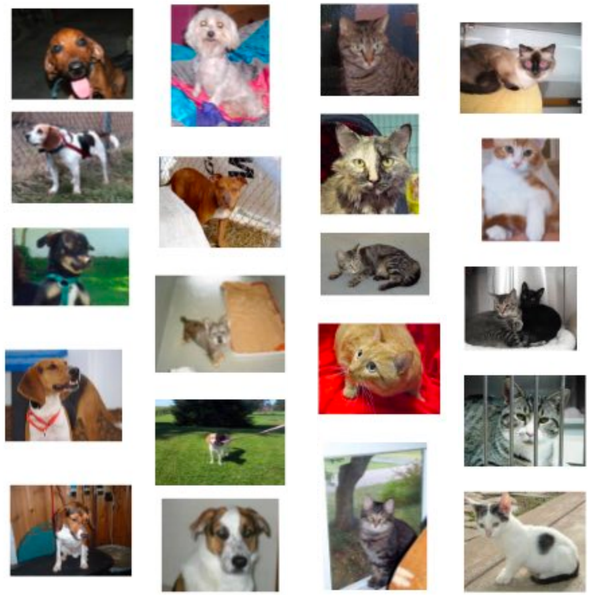


Gatti e Cani

TEST

Gatti e Cani

VALIDAZIONE

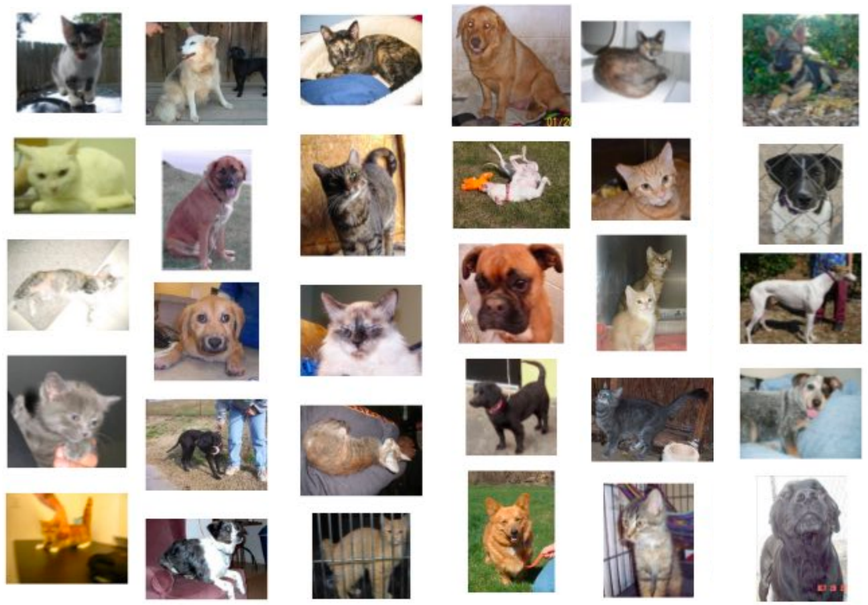


Gatti e Cani

ELABORAZIONE DELLA TEORIA E VALIDAZIONE

Classificatore di **cani** e **gatti**

TRAIN



Gatti e Cani

Qui l'algoritmo ha a disposizione sia le immagini che l'etichetta CANE/GATTO **la macchina impara** (fissa i parametri)

Qui l'algoritmo ha a disposizione solo le immagini **posso verificare** se la macchina ha imparato

Se non ha imparato "**aggiusto**" il modello per migliorare la predittività

TEST

Gatti e Cani

ELABORAZIONE DELLA TEORIA E VALIDAZIONE

Classificatore di **cani** e **gatti**

Utilizzo queste immagini sul modello ottimizzato nella fase di training e test su immagini mai viste prima

Se il modello funziona anche qui sono confidente che possa funzionare su qualsiasi nuova immagine

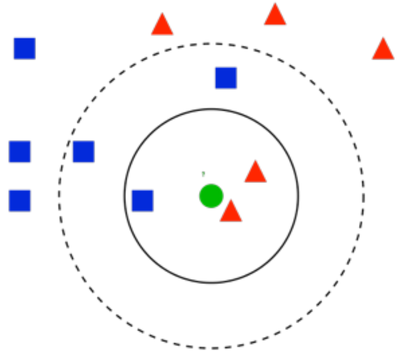
FONDAMENTALE PER UN APPROCCIO **SCIENTIFICO** ALLA COSTRUZIONE DI MODELLI CON **MACHINE LEARNING**

QUALE ALGORITMO?

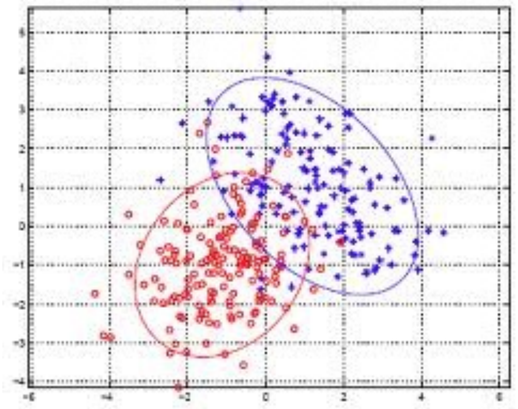
Una delle cose più difficili ed importanti è scegliere l'algoritmo giusto (o inventarne uno nuovo)

CLASSIFICATORI

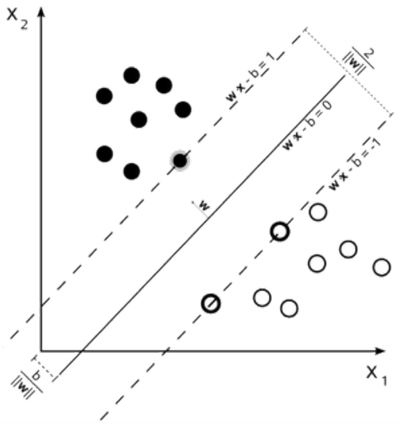
k-nearest neighbour



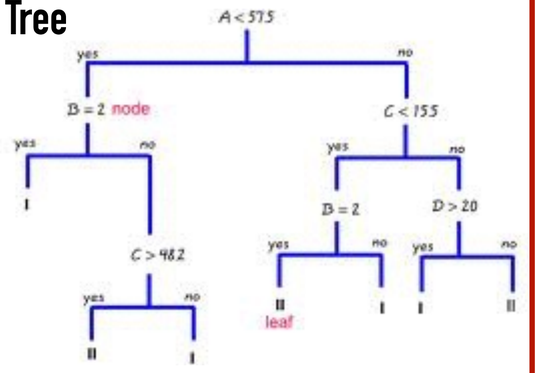
Linear Discriminant Analysis



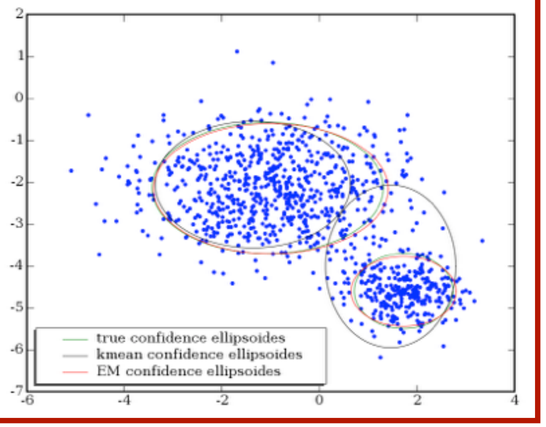
Support Vector Machine



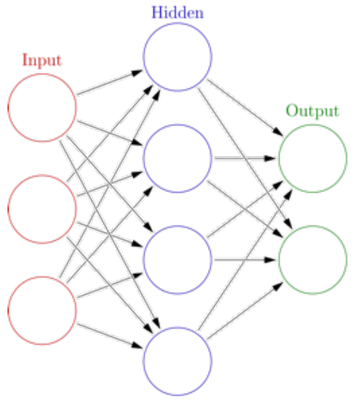
Decision Tree



Gaussian Mixture



Artificial Neural Network (Deep Learning)



QUALE ALGORITMO? Es. IRIS DATASET [Fisher 1936]

× Iris setosa

○ Iris virginica

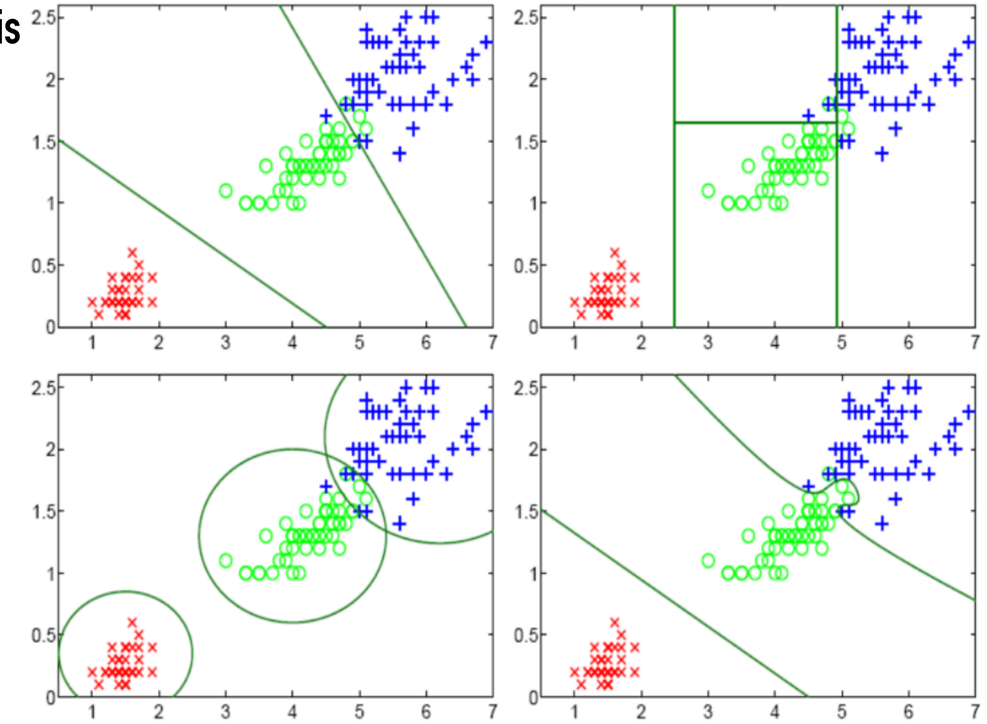
\+ Iris versicolor

50 campioni per ogni tipo di fiore

4 caratteristiche: lunghezza e larghezza di sepali e petali

Linear Discriminant Analysis

Decision Tree



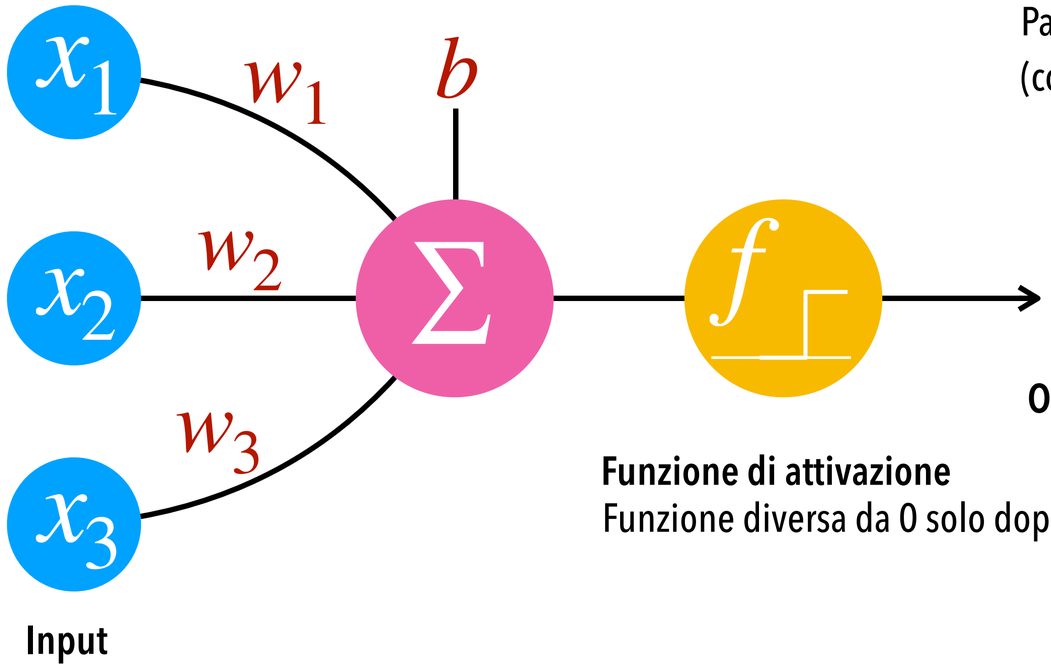
Gaussian Mixture

Support Vector Machine

NEURAL NETWORKS & DEEP LEARNING

Nuova speranza per AI e boom del Machine Learning nelle scienze dure

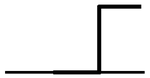
Neurone



w_1 w_2 w_3 b

Parametri che la macchina deve imparare
(come nell'esempio dei girasoli)

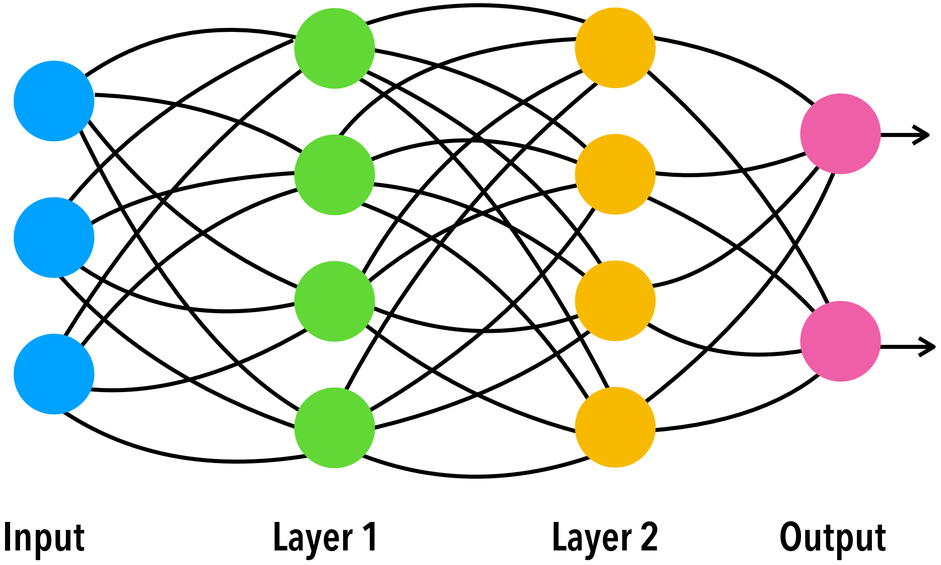
Funzione di attivazione
Funzione diversa da 0 solo dopo una soglia



$$y = f\left(b + \sum_i x_i w_i\right)$$

NEURAL NETWORKS & DEEP LEARNING

Nuova speranza per AI e boom del Machine Learning nelle scienze dure



Aggiungendo layer la capacità di imparare della macchina aumenta

Aggiungendo layer diventa sempre più difficile capire che cosa la macchina estrae dai dati per imparare

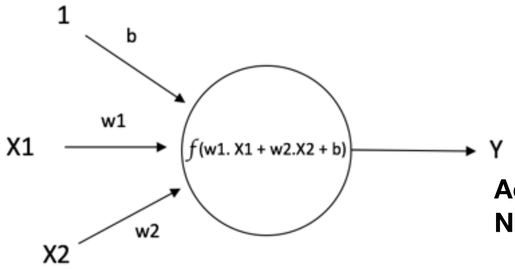
BLACK BOX

NEURAL NETWORKS & DEEP LEARNING

Nuova speranza per AI e boom del Machine Learning nelle scienze dure

Ingredienti

Neuron



Linear part

$$f(w_1 \cdot x_1 + w_2 \cdot x_2 + b)$$

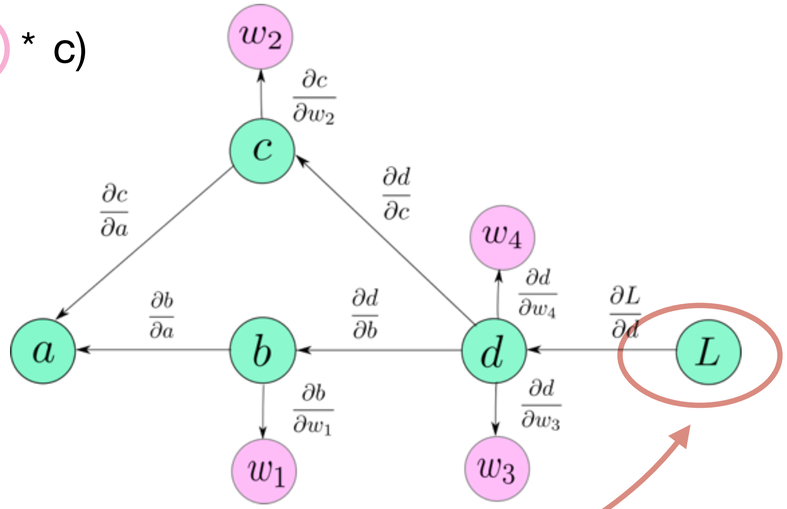
Activation function
Non linearity

Automatic differentiation (chain rule)

b = w1 * a
c = w2 * a
d = (w3 * b) + (w4 * c)
L = f(d)

weights

→ **Backpropagation**



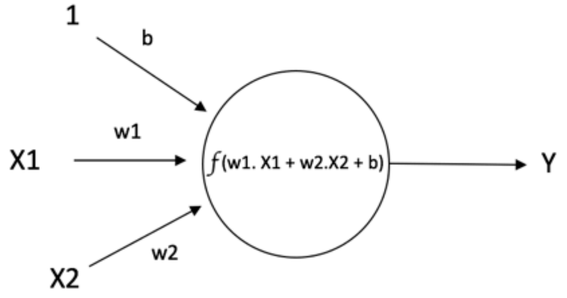
Loss function (a scalar to minimise)

- Cross Entropy for classification
- Mean Square Error for regression
- Energy for physical problems?

NEURAL NETWORKS & DEEP LEARNING

Nuova speranza per AI e boom del Machine Learning nelle scienze dure

The neuron

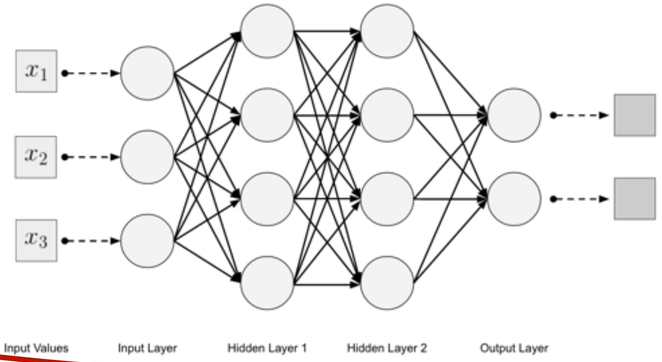


Linear part

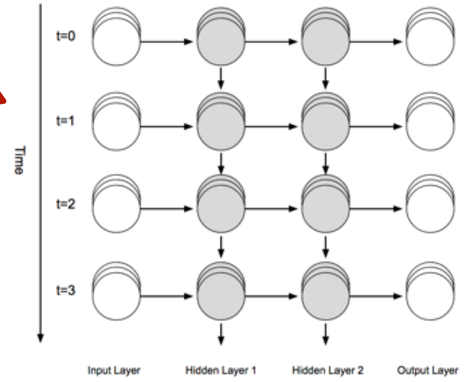
$$f(w_1 \cdot x_1 + w_2 \cdot x_2 + b)$$

Activation function
Non linearity

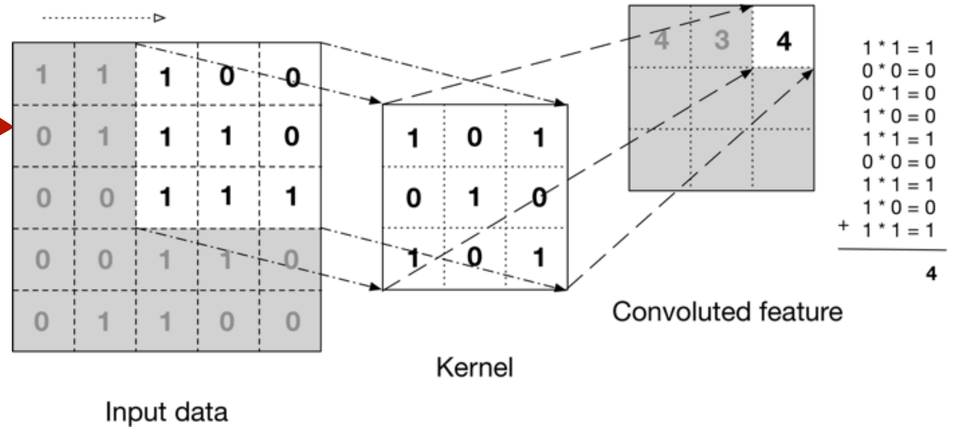
Fully connected Network



Recurrent Network



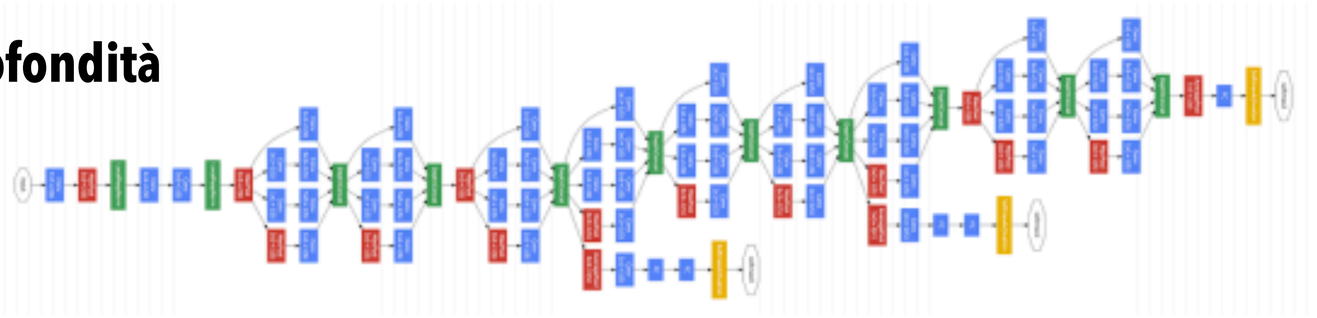
Convolutional Network



NEURAL NETWORKS & DEEP LEARNING

Nuova speranza per AI e boom del Machine Learning nelle scienze dure

Si può scendere molto in profondità



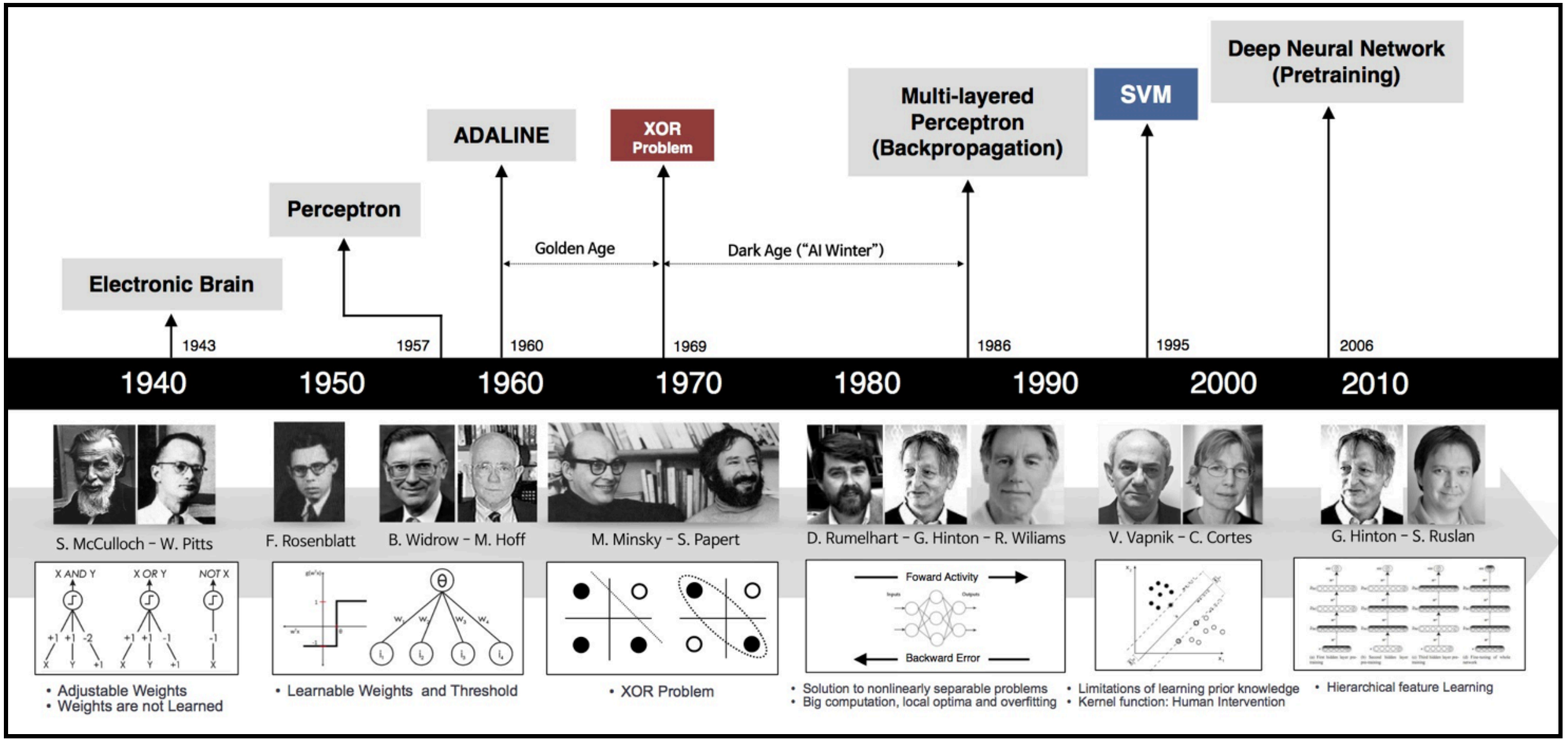
[1]Inception 5 (GoogLeNet)

Inception 7a

[1]Going Deeper with Convolutions, [C. Szegedy et al, CVPR 2015]

NEURAL NETWORKS & DEEP LEARNING

Nuova speranza per AI e boom del Machine Learning nelle scienze dure



FONDAZIONE BRUNO KESSLER

MPBA PREDICTIVE MODELS FOR BIOMEDICINE & ENVIRONMENT

TIFPA Trento Institute for Fundamental Physics and Applications

NEURAL NETWORKS & DEEP LEARNING

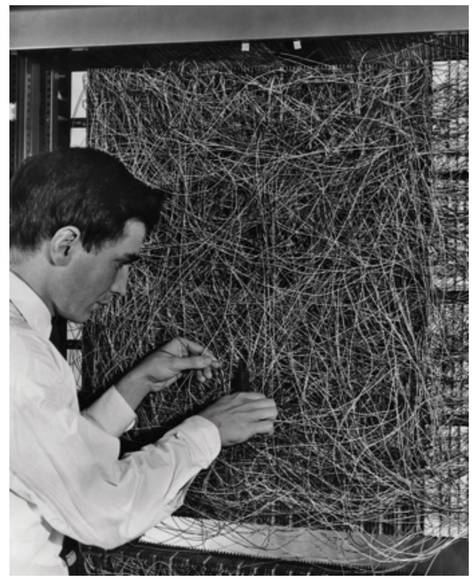
Nuova speranza per AI e boom del Machine Learning nelle scienze dure

Due ingredienti fondamentali per la rinascita

Disponibilità di enormi quantità di dati per allenare i modelli

Risorse di calcolo inimmaginabili fino a qualche anno prima [GPU]

The Mark I Perceptron - 1960



FONDAZIONE BRUNO KESSLER

MPBA PREDICTIVE MODELS FOR BIOMEDICINE & ENVIRONMENT

TIFPA Trento Institute for Fundamental Physics and Applications

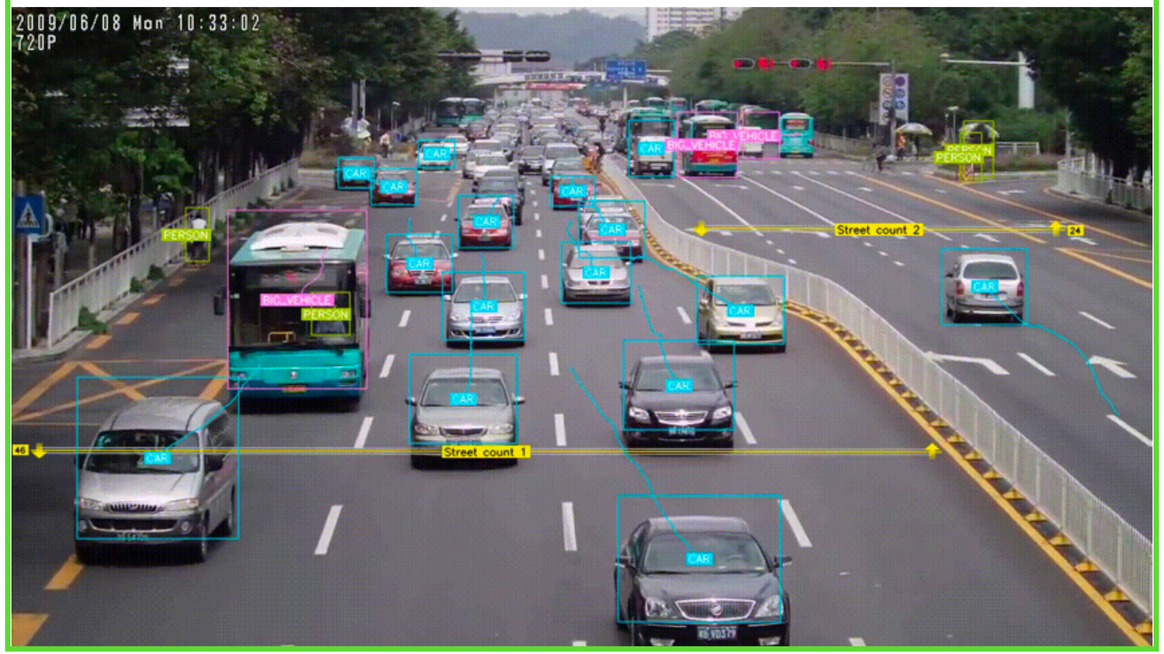
NEURAL NETWORKS & DEEP LEARNING

Nuova speranza per AI e boom del Machine Learning nelle scienze dure

2016: AlphaGo@DeepMind

TECNOLOGIA e AI

OBJECT DETECTION



ASSISTENTE VIRTUALE

Google Duplex

Advancing AI for Everyone

NEURAL NETWORKS & DEEP LEARNING

Nuova speranza per AI e boom del Machine Learning nelle scienze dure

SALUTE

Oftalmologia

Press Release: FDA permits marketing of IDx-DR for automated detection of diabetic retinopathy in primary care

April 12, 2018

2018

Tumori della pelle

Tumore al fegato

Published OnlineFirst October 5, 2017; DOI: 10.1158/1078-0432.CCR-17-0853

Statistics in CCR

Clinical Cancer Research

Deep Learning–Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer

Kumardeep Chaudhary[1], Olivier B. Poirion[1], Liangqun Lu[1,2], and Lana X. Garmire[1,2]

Abstract

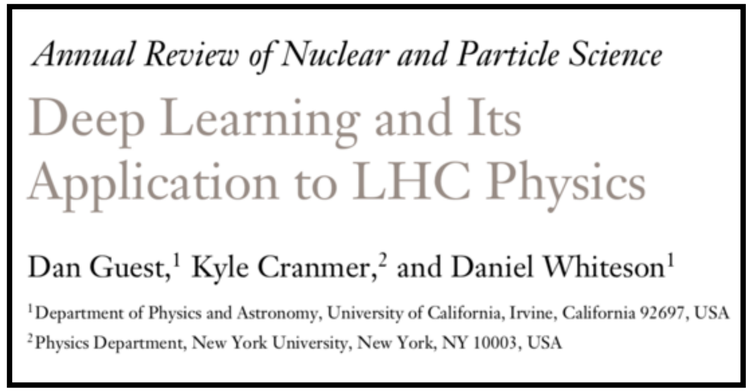
Identifying robust survival subgroups of hepatocellular carcinoma (HCC) will significantly improve patient care. Currently, endeavor of integrating multi-omics data to explicitly predict HCC survival from multiple patient cohorts is lacking. To fill this gap, we present a deep learning (DL)–based model on HCC that robustly differentiates survival subpopulations of patients in six cohorts. We built the DL-based, survival-sensitive model on 360 HCC patients' data using RNA sequencing (RNA-Seq), miRNA sequencing (miRNA-Seq), and methylation data from The Cancer Genome Atlas (TCGA), which predicts prognosis as good as an alternative model where genomics and clinical data are both considered. This DL-based model provides two optimal subgroups of patients with significant survival differences ($P = 7.13e{-}6$) and good model fitness [concordance index (C-index) = 0.68]. More aggressive subtype is associated with frequent *TP53* inactivation mutations, higher expression of stemness markers (*KRT19* and *EPCAM*) and tumor marker *BIRC5*, and activated Wnt and Akt signaling pathways. We validated this multi-omics model on five external datasets of various omics types: LIRI-JP cohort ($n = 230$, C-index = 0.75), NCI cohort ($n = 221$, C-index = 0.67), Chinese cohort ($n = 166$, C-index = 0.69), E-TABM-36 cohort ($n = 40$, C-index = 0.77), and Hawaiian cohort ($n = 27$, C-index = 0.82). This is the first study to employ DL to identify multi-omics features linked to the differential survival of patients with HCC. Given its robustness over multiple cohorts, we expect this workflow to be useful at predicting HCC prognosis prediction. *Clin Cancer Res; 24(6); 1248–59. ©2017 AACR.*

2018

NEURAL NETWORKS & DEEP LEARNING

Nuova speranza per AI e boom del Machine Learning nelle scienze dure

SCIENZE DURE



Annual Review of Nuclear and Particle Science

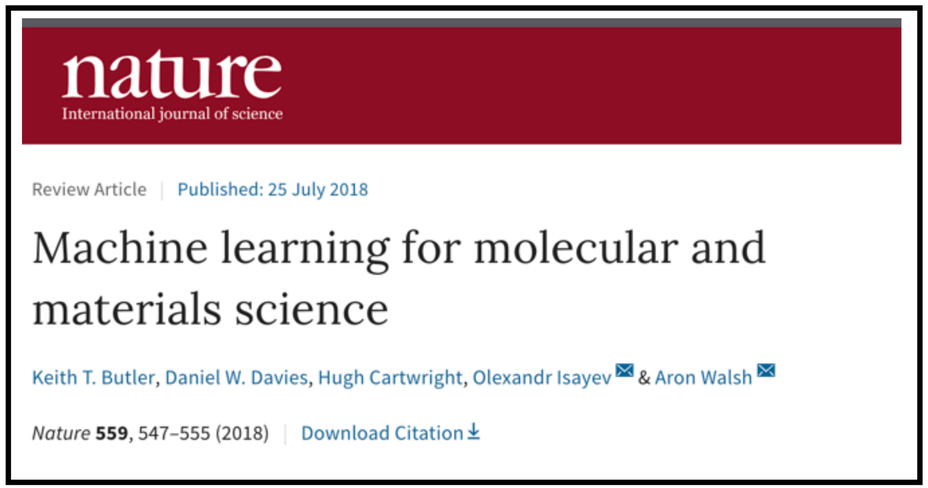
Deep Learning and Its Application to LHC Physics

Dan Guest,[1] Kyle Cranmer,[2] and Daniel Whiteson[1]

[1]Department of Physics and Astronomy, University of California, Irvine, California 92697, USA

[2]Physics Department, New York University, New York, NY 10003, USA

Fisica



nature

International journal of science

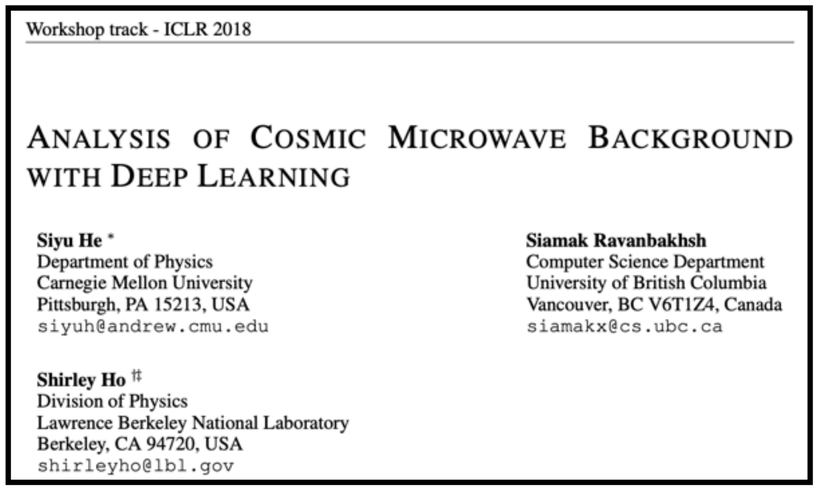
Review Article | Published: 25 July 2018

Machine learning for molecular and materials science

Keith T. Butler, Daniel W. Davies, Hugh Cartwright, Olexandr Isayev & Aron Walsh

Nature **559**, 547–555 (2018) | Download Citation

Chimica / Scienze dei materiali



Workshop track - ICLR 2018

ANALYSIS OF COSMIC MICROWAVE BACKGROUND WITH DEEP LEARNING

Siyu He *
Department of Physics
Carnegie Mellon University
Pittsburgh, PA 15213, USA
siyuh@andrew.cmu.edu

Siamak Ravanbakhsh
Computer Science Department
University of British Columbia
Vancouver, BC V6T1Z4, Canada
siamakx@cs.ubc.ca

Shirley Ho ♯
Division of Physics
Lawrence Berkeley National Laboratory
Berkeley, CA 94720, USA
shirleyho@lbl.gov

Astrofisica

COME SI DIVENTA DATA SCIENTIST

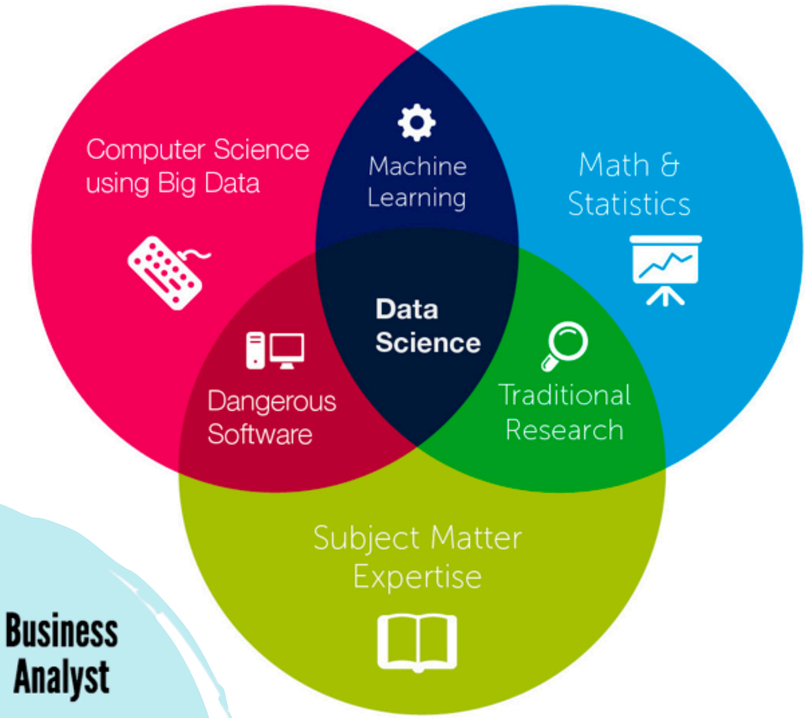
Disciplina nuova

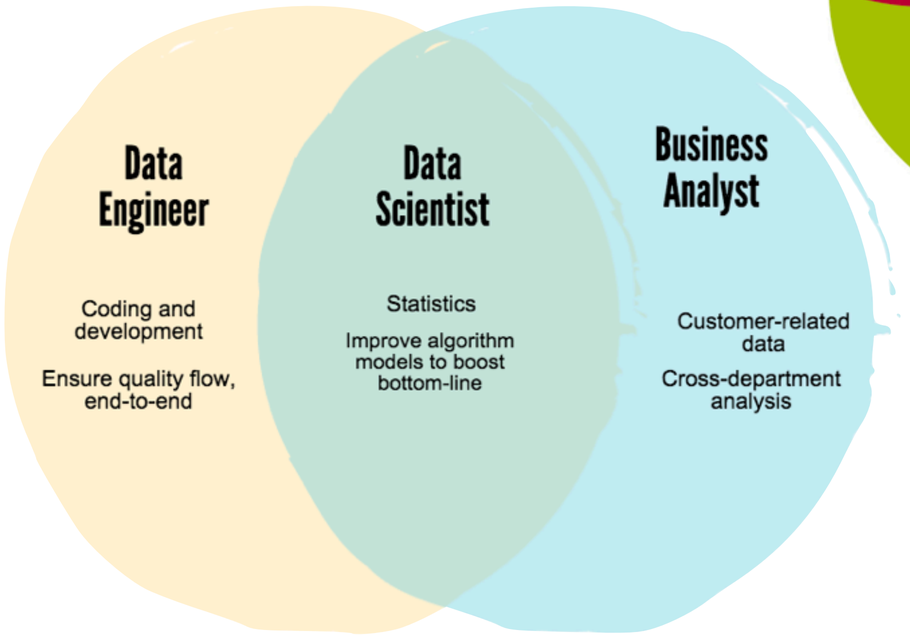
recentemente le Università hanno iniziato a proporre corsi di laurea in Data Science

Competenze: **Computer science**
Matematica
Statistica

ma anche **Scienze sociali**

Sviluppo soft skills fondamentale





FONDAZIONE BRUNO KESSLER

MPBA PREDICTIVE MODELS FOR BIOMEDICINE & ENVIRONMENT

TIFPA Trento Institute for Fundamental Physics and Applications

DATA SCIENTIST A 17 ANNI

Cesare Furlanello, Claudia Dolci, Giuseppe Jurman

Camp di Data Science: 20 studenti (17 - 19 anni) internazionali
per 3 settimane di full immersion nella ricerca sulle montagne del Trentino
alla scoperta di soluzioni innovative per challenge sul mondo reale

FONDAZIONE BRUNO KESSLER

MPBA PREDICTIVE MODELS FOR BIOMEDICINE & ENVIRONMENT

TIFPA Trento Institute for Fundamental Physics and Applications

Grazie