D. *Trinchese* - C. *Napolitano* - A. *Gravino*

Strumenti Formali per la Bioinformatica

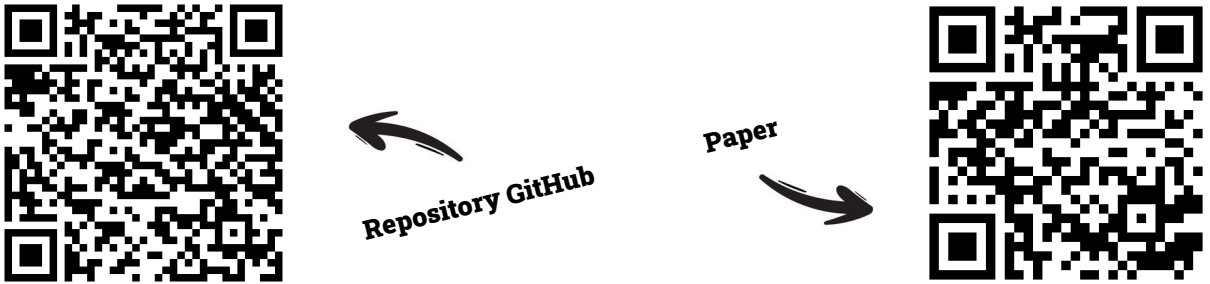
Digital Twin e le Opportunità del Cloud

Dipartimento di Informatica
Università degli Studi di Salerno
A.A 2022/23

Il nostro progetto

Sarà analizzato in dettaglio un esempio di applicazione in ambito medico di un sistema basato su tecnologia "Digital Twin", descrivendone i principi fondamentali, tecnologie necessarie per l'implementazione, modifiche apportate in seguito ad un lavoro di analisi del sistema.

Per tale applicazione sarà presentato inoltre una proof-of-concept di utilizzo di tecnologie di cloud-computing a supporto aggiuntivo delle operazioni.



Punti chiave - concetti fondamentali (qui va la scaletta)

1. Digital Twin.
   a. Cosa sono?
   b. Benefici derivanti dall'uso di DTs
   c. Domini applicativi dei DTs
   d. Esempi concreti di applicazioni DT-based
   e. Opportunità sul cloud: il nostro progetto, obiettivi, motivazioni
2. La sperimentazione
   a. Alla base del Patient's DT : GNN
   b. Scenario clinico
   c. Aspetti implementativi
   d. Risultati
3. Bioinformatica sul Cloud
   a. Perché il Cloud: il caso di AWS
   b. Containerizzare un progetto di Bioinformatica con Docker
   c. Creare una routine di CI/CD per un progetto Python di ML
   d. Interagire con il modello di ML in Flask
   e. Sviluppi futuri

1

Digital Twins

Cosa sono? Benefici, domini applicativi ed esempi concreti

Opportunità sul Cloud

Obiettivi e motivazioni

Digital Twin: cos'è?

Modello digitale di un **sistema fisico** e dei suoi processi in corso

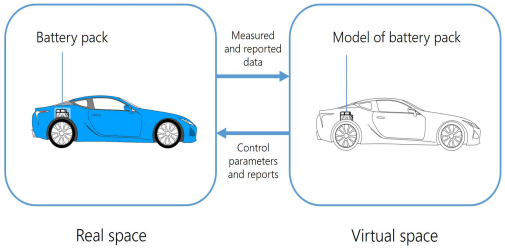
In termini più semplici? **Esatta replica virtuale di un modello fisico.**

DT ideale → modello digitale e fisico sono *altamente sincronizzati*

- Sincronia ottenuta tramite flusso di dati *bidirezionale*
- Dati prelevati tramite l'uso di *sensori*

Qualunque sistema fisico può essere replicato digitalmente!

- Un auto
- Un'azienda
- Un ponte
- ...
- Un **paziente**! (ci torneremo dopo)



Esempio di modellazione: Digital Twin per modellare il battery-pack di un automobile

“Il Digital Twin non è niente di nuovo”

through its operational life, and is eventually retired and disposed of.

In the create phase, the physical system does not yet exist. The system starts to take shape in virtual space as a Digital Twin Prototype (DTP). This is not a new phenomenon. For most of human history, the virtual space where this system was created existed only in people’s minds. It is only in the last quarter of the 20th century that this virtual space could exist within the digital space of computers.

This opened up an entire new way of system creation. Prior to this leap in technology, the system would have to have been implemented in physical form, initially in sketches and blueprints but shortly thereafter made into costly prototypes, because simply existing in people’s minds meant very limited group sharing and understanding of both form and behavior.

In addition, while human minds are a marvel, they have severe limitations

Capitolo: The Digital Twin Concept
Sezione “Defining the Digital Twin”

Origins of the Digital Twin Concept

August 2016

DOI:10.13140/RG.2.2.26367.61609

Affiliation: Florida Institute of Technology / NASA

Project: Digital Twin

Authors:

Michael Grieves
Digital Twin Institute

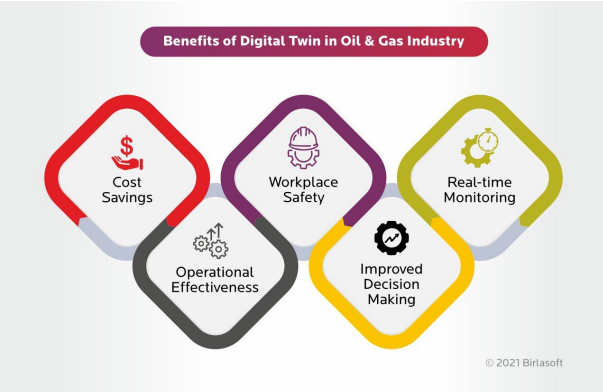
Digital Twin: Mitigating Unpredictable, Undesirable Emergent Behavior in Complex Systems (Excerpt)

Benefici derivanti dall'uso di tecnologie DT-based

Utilizzare tecnologie Digital-Twin-based porta a **notevoli benefici**!

In nessun ordine particolare, alcuni di questi possono essere:

- **Monitoraggio in tempo reale, controllo e acquisizione dei dati**
  - DT e corrispondente sistema fisico sono equivalenti
  - Analizzare del DT piuttosto che il sistema reale
- **Supporto alle decisioni**
  - Decisioni aziendali potrebbero dipendere dal sistema fisico
  - Analizzare il DT può sicuramente velocizzare il processo
- **Risk assessment migliorato**
  - DT e corrispondente sistema fisico, ancora, sono equivalenti
  - Testare sul DT piuttosto che sul sistema fisico!
- **Efficienza aumentata**
  - Deriva dai vantaggi precedenti



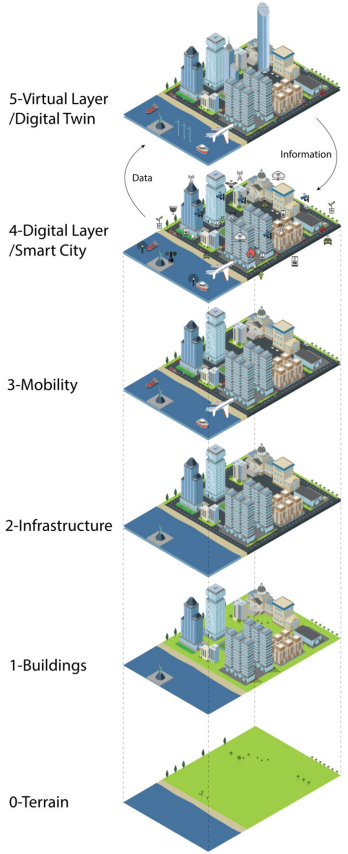
Benefici dell'uso di **DTs** nel settore industriale **del Gas** e del **Petrolio** (2021 – Birlasoft)

Digital Twins: potenziali domini applicativi

Grazie ai numerosi vantaggi descritti, le tecnologie DT-based sembrano essere particolarmente attraenti ed utili per diversi scopi…

Hanno trovato presto applicazione in diversi domini:

- **Settore industriale e manifatturiero**
  - DTs delle linee di produzione (o di intere fabbriche) su cui possono essere testati ed ottimizzati tutti i processi.
- **Smart cities**
  - Visualizzazione delle risorse della città, monitoraggio delle infrastrutture, delle attività commerciali, nonché pianificazione di sviluppi futuri
- **Istruzione**
  - ambienti di apprendimento intelligenti con i framework di apprendimento adattativo personalizzati



Schema di riferimento per la modellazione del DT di una città

“L’utilizzo di tecnologie basate sull’uso di Digital Twin potrebbero avere grandi risvolti in ambito medico”

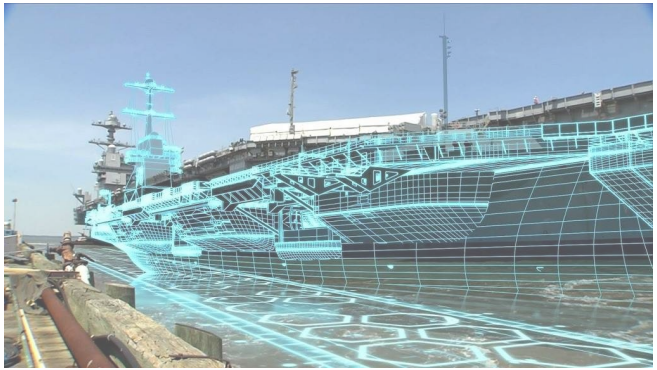
Digital Twin for Healthcare

Il modello dei Digital Twin ha la potenzialità di *rivoluzionare l’assistenza sanitaria* e permettere **terapie personalizzate** in base alle **caratteristiche** e ai **bisogni** di ciascun paziente.

> Creare un gemello digitale di una persona, inoltre, permetterebbe al personale sanitario di avere un **quadro completo del paziente**, *indipendentemente dalla struttura* e dal *personale* a cui si rivolge.

Creare una replica digitale di un paziente, con tutte le sue informazioni – *dna*, *patologie e particolari predisposizioni*, *storia familiare ecc.* – permetterà in futuro agli specialisti sanitari di scegliere **la migliore terapia per il singolo**.

Esempio concreto: AWARE



FINCANTIERI si è classificata tra le aziende vincitrici del secondo bando del *Competence Center Smact* con un progetto finalizzato alla **creazione di un gemello digitale** per *ottimizzare e ridurre i costi del processo di costruzione di una nave*.

Nel progetto **AWARE** queste tecnologie (DT e IoT) saranno adattate a due ambienti manifatturieri con differenti livelli di complessità: un cantiere navale e uno stabilimento di produzione macchinari.

1. Il primo fornisce la possibilità di *simulare le attività quotidiane* e *fornire informazioni critiche sul funzionamento dei processi in esecuzione* e *tenere sotto controllo lo svolgimento di tutte le attività produttive*

2. Il secondo sarà invece applicato ad un *macchinario per la produzione e il controllo di pacchi statorici o rotorici di motori elettrici*.

Digital Twin in ambito biomedico? Diversi esempi

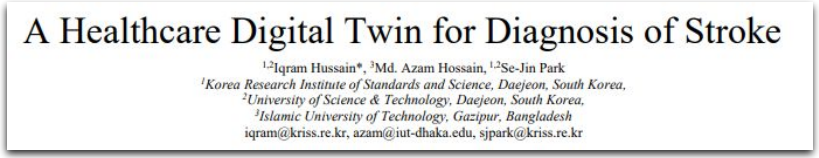
Le proposte di applicazioni DT-based in ambito biomedico sono *numerose* (ne ho riportate giusto alcune…)

L'uso dei Digital Twins al servizio dell'healthcare è **affascinante tanto quanto sembra**!

Dall'uso di DTs per realizzare il **gemello digitale di un paziente** al coinvolgimento dei **DT per le strutture sanitarie** (*ospedali*, *reparti*, *pronto soccorso*, etc…)


Digital Twins in Healthcare: an architectural proposal and its application in a social distancing case study

Alessandra De Benedictis, Nicola Mazzocca, Alessandra Somma, and Carmine Strigaro


A Healthcare Digital Twin for Diagnosis of Stroke

[1,2]Iqram Hussain*, [3]Md. Azam Hossain, [1,2]Se-Jin Park
[1]Korea Research Institute of Standards and Science, Daejeon, South Korea,
[2]University of Science & Technology, Daejeon, South Korea,
[3]Islamic University of Technology, Gazipur, Bangladesh
iqram@kriss.re.kr, azam@iut-dhaka.edu, sjpark@kriss.re.kr


Human Digital Twin for Personalized Healthcare: Vision, Architecture and Future Directions

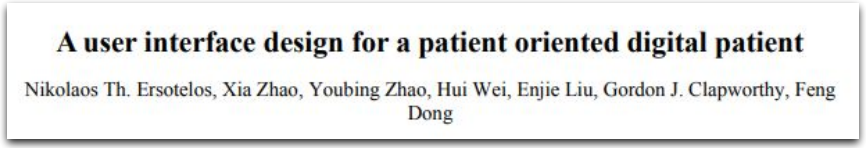
Samuel D. Okegbile, *Member, IEEE*, Jun Cai, *Senior Member, IEEE*, Changyan Yi, *Member, IEEE*, and Dusit Niyato, *Fellow, IEEE*


The Digital Twin Revolution in Healthcare

Tolga Erol
Training and Simulation Technologies
HAVELSAN A.Ş.
Ankara, Turkey
terol@havelsan.com.tr

Arif Furkan Mendi
Training and Simulation Technologies
HAVELSAN A.Ş.
Ankara, Turkey
afmendi@havelsan.com.tr

Dilara Doğan
Training and Simulation Technologies
HAVELSAN A.Ş
Ankara, Turkey
ddogan@havelsan.com.tr


A user interface design for a patient oriented digital patient

Nikolaos Th. Ersotelos, Xia Zhao, Youbing Zhao, Hui Wei, Enjie Liu, Gordon J. Clapworthy, Feng Dong

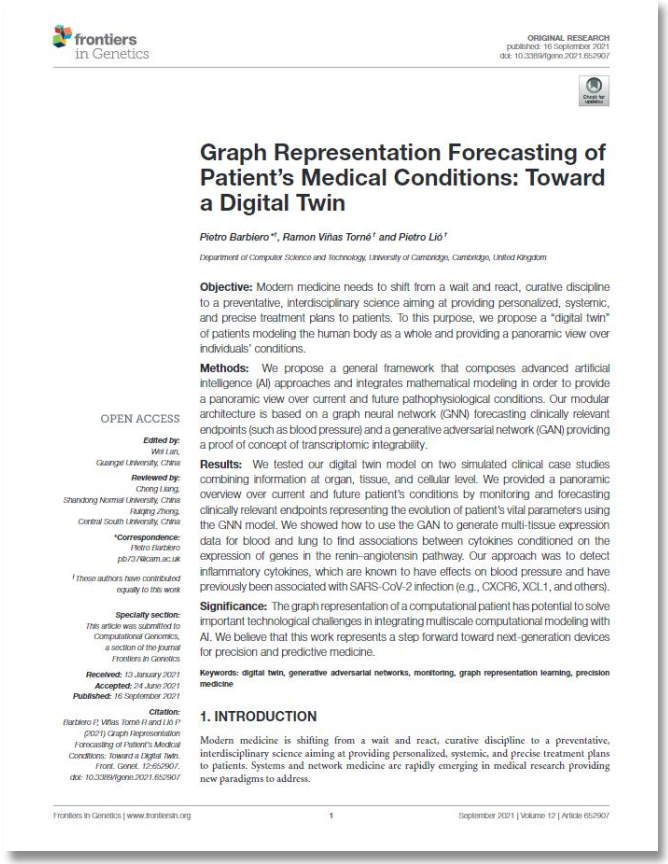
Altro esempio: Patient's Digital Twin [1]

Un esempio concreto e molto interessante di un'**applicazione basata su tecnologie Digital-Twin in ambito medico**.

È il lavoro di riferimento per lo sviluppo del nostro progetto!

In **"Graph Representation Forecasting of Patient's Medical Conditions: Toward a Digital Twin"** (*Pietro Barbiero et al.*) gli autori propongono un gemello digitale per pazienti in grado di *modellare il corpo umano* e fornisce una *visione panoramica delle sue caratteristiche*.

Gli autori, nell'introduzione della pubblicazione, ci tengono a **rimarcare l'obiettivo della medicina moderna**: traslare dall'essere una *disciplina di wait and react* ad una *scienza interdisciplinare.*



frontiers in Genetics

ORIGINAL RESEARCH
published: 16 September 2021
doi: 10.3389/fgene.2021.652907

Graph Representation Forecasting of Patient's Medical Conditions: Toward a Digital Twin

Pietro Barbiero*†, Ramon Viñas Torné† and Pietro Lió†

Department of Computer Science and Technology, University of Cambridge, Cambridge, United Kingdom

Objective: Modern medicine needs to shift from a wait and react, curative discipline to a preventative, interdisciplinary science aiming at providing personalized, systemic, and precise treatment plans to patients. To this purpose, we propose a "digital twin" of patients modeling the human body as a whole and providing a panoramic view over individuals' conditions.

Methods: We propose a general framework that composes advanced artificial intelligence (AI) approaches and integrates mathematical modeling in order to provide a panoramic view over current and future pathophysiological conditions. Our modular architecture is based on a graph neural network (GNN) forecasting clinically relevant endpoints (such as blood pressure) and a generative adversarial network (GAN) providing a proof of concept of transcriptomic integrability.

Results: We tested our digital twin model on two simulated clinical case studies combining information at organ, tissue, and cellular level. We provided a panoramic overview over current and future patient's conditions by monitoring and forecasting clinically relevant endpoints representing the evolution of patient's vital parameters using the GNN model. We showed how to use the GAN to generate multi-tissue expression data for blood and lung to find associations between cytokines conditioned on the expression of genes in the renin-angiotensin pathway. Our approach was to detect inflammatory cytokines, which are known to have effects on blood pressure and have previously been associated with SARS-CoV-2 infection (e.g., CXCR6, XCL1, and others).

Significance: The graph representation of a computational patient has potential to solve important technological challenges in integrating multiscale computational modeling with AI. We believe that this work represents a step forward toward next-generation devices for precision and predictive medicine.

Keywords: digital twin, generative adversarial networks, monitoring, graph representation learning, precision medicine

OPEN ACCESS

Edited by: Wei Lan, Guangxi University, China

Reviewed by: Cheng Liang, Shandong Normal University, China; Huiqing Zhong, Central South University, China

***Correspondence:** Pietro Barbiero pb737@cam.ac.uk

† These authors have contributed equally to this work

Specialty section: This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: 13 January 2021
Accepted: 24 June 2021
Published: 16 September 2021

Citation: Barbiero P, Viñas Torné R and Lió P (2021) Graph Representation Forecasting of Patient's Medical Conditions: Toward a Digital Twin. Front. Genet. 12:652907. doi: 10.3389/fgene.2021.652907

1. INTRODUCTION

Modern medicine is shifting from a wait and react, curative discipline to a preventative, interdisciplinary science aiming at providing personalized, systemic, and precise treatment plans to patients. Systems and network medicine are rapidly emerging in medical research providing new paradigms to address.

"Graph Representation Forecasting of Patient's Medical Conditions: Toward a Digital Twin" (*Pietro Barbiero et al.*)

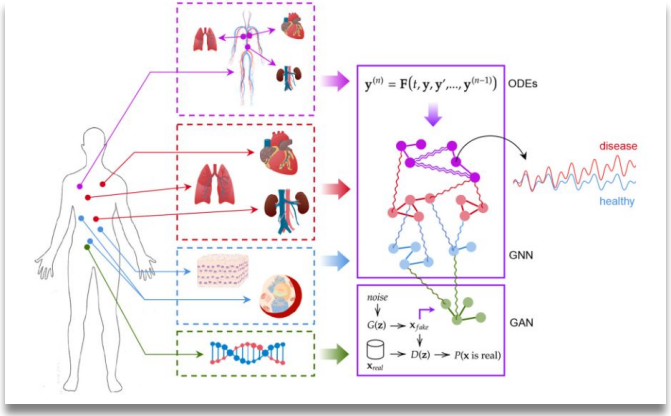
Patient's DT: dettagli [1]

L'applicazione oggetto di studio di questo lavoro è disponibile su *un'apposita repository GitHub* (***vedi QR-code***).

Il Digital Twin realizzato (da Pietro Barbiero et al.) consiste in un sistema di intelligenza artificiale modulare che può essere utilizzato per **modellare il corpo umano** nel suo insieme e per **prevedere l'evoluzione delle condizioni patofisiologiche**.

Si compone di due moduli

- Il primo modulo si basa su una rete neurale a grafo (**GNN**) che *prevede endpoint clinicamente rilevanti* (come la **pressione sanguigna** e dati relativi al **sistema renina-angiotensina**)
- Il secondo invece è rappresentato da una rete generativa avversaria (**GAN**) che fornisce una proof-of-concept di integrabilità multi-omica.



Modello del Digital Twin di (*Pietro Barbiero et al.*)

https://github.com/pietrobarbiero/digital-patient

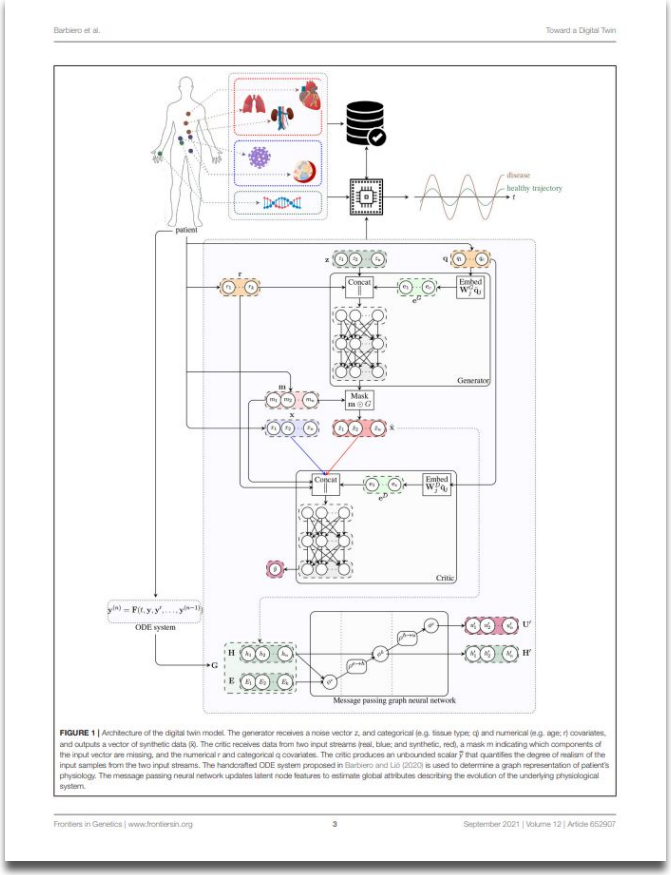
Patient's DT: dettagli [2]

Il nostro progetto pone attenzione sul primo modulo, che fa uso di una **Graph-Neural-Network** (**GNN**) per il **forecasting di endpoint sanitari rilevanti**. Nell'applicazione sviluppata da Barbiero et al. viene utilizzata una GNN per la modellazione della complessità del corpo umano.

La complessità del corpo umano è in primo luogo scomposta in sottosistemi indipendenti. Ogni sottosistema è rappresentato nella GNN da *un nodo* o da *un network di nodi*. Sottosistemi omogenei ed affini possono essere aggregati in **layer**.

Il loro modello di Digital Patient comprende **quattro** layer:

- *Transcriptomic layer*
- *Cellular layer*
- *Organ layer*
- *Exposomic layer*



Architettura del Digital Twin Model (*Pietro Barbiero et al.*)

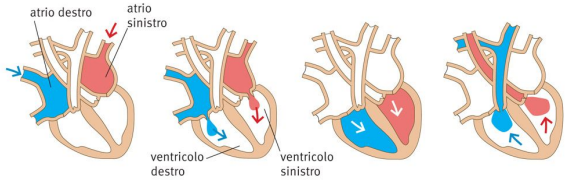
Patient's DT: dettagli [3]

Transcriptomic layer: opera sull'*insieme dei trascritti prodotti dal genoma in un certo istante*. Nel modello viene studiata la comunicazione tra tessuti nell'organ-layer attraverso il ***communicome*** (*fattori di comunicazione nel sangue*).

Cellular Layer: il layer cellulare modella il **RAS** (**Sistema Renina-Angiotensina**)

- **RAS**: sistema ormonale che regola la vasocostrizione e la risposta infiammatoria
- L'ormone chiave è **ANG-II**, generato dall'enzima di conversione **ACE**
- **SARS-CoV-2** si lega all'enzima **ACE2** per entrare nella cellula ospite
- *SARS-CoV-2 nel paziente ⟹ Possibili alterazioni della concentrazione di ACE2*

Organ layer: modella gli organi e network di organi cooperanti. È limitato alla rappresentazione del sistema circolatorio e di alcuni organi (cuore, polmoni).



Ciclo cardiaco

Il modello del cuore include : *Atrio Sinistro*, *Atrio Destro*, *Ventricolo Sinistro*, *Ventricolo Destro*.

Exposomic layer: si riferisce alle esposizioni che gli individui sperimentano dal concepimento fino alla morte, Regimi alimentari, esercizio fisico, postura, abitudini di vita ed esposizione a sostanze tossiche sono possibili esposizioni che contribuiscono al benessere o alle condizioni di malattia dell'individuo.

Opportunità sul cloud

Il nostro progetto

Presentare una proof-of-concept per un sistema che, oltre a fornire **strumenti di machine-learning** che operano nel **campo della medicina**, integra al suo interno **elementi di tecnologie cloud**.

Questo obiettivo fonda le sue motivazioni nei **vantaggi** intrinseci che deriverebbero dall'**uso di Cloud-Computing technologies** in ambito Machine-Learning.

Le tecnologie cloud offrono la capacità di *scalare rapidamente* e *in modo efficace* le applicazioni di machine learning che richiedono *maggiore capacità di elaborazione*.

Opportunità sul cloud

Il nostro progetto

Perché questa direzione?

Progressi nella medicina e nell'informatica
⇓
Necessità di paradigmi moderni per le soluzioni

I progressi nella medicina e nell'informatica possano, in un futuro molto prossimo, richiedere **notevoli quantità di potenza computazionale** così come grandi necessità di **scalabilità**, **affidabilità** e **sicurezza** sicuramente desiderabili in un campo critico quanto quello generato dalle *sinergie tra medicina ed informatica*.

2

Sperimentazione

1. Alla base del Patient's DT: perché le GNN?
2. Scenario clinico
3. Aspetti implementativi
4. Risultati

Graph Neural Network [1]

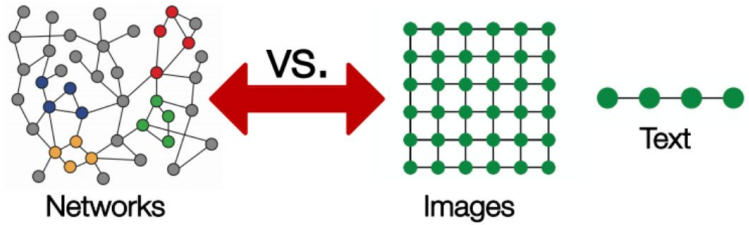
Gli oggetti del mondo reale sono spesso definiti in termini di **connessioni**. Un insieme di oggetti e le connessioni tra di essi sono naturalmente espressi come un **grafo**.

Un **grafo** G può essere definito come **G = (V, E)**, dove **V** è l'insieme dei nodi, ed **E** l'insieme degli archi

Una **GNN** rappresenta un modello di deep learning che lavora sul dominio dei grafi. Alla base delle GNN vi sono dunque grafi e reti neurali.Dalle reti neurali le GNN ereditano un approccio basato sui dati associato ad un'architettura multilivello, dai grafi invece derivano un'estrema flessibilità ed interpretabilità.

Perché le GNN?

Gli strumenti tradizionali di Machine Learning e Deep Learning, come ad esempio le Convolutional Neural Network (CNN), sono specializzati nell'elaborazione di dati semplici. Esempi di dati semplici possono essere le immagini che possiamo vedere come dei grafi a griglia, oppure il testo che possiamo vedere come dei grafi orientati in cui ogni nodo (parola) è connesso al nodo seguente (parola seguente)



Le GNN invece potendo essere applicate direttamente ai grafi, forniscono un modo semplice per fare attività di previsione sia a livello di nodi, sia a livello di archi, sia a livello di grafo.

Graph Neural Network [2]

Framework Battaglia et al.

Tale framework è basato sulle GNN, dove vi sono i graph network block (GN blocks) che rappresentano le unità di calcolo fondamentali della GNN.

Definisce una GNN come una tupla G=(u, H, E). Dove:

H=$\{h_i\}_{i=1:N_v}$ è l'insieme dei nodi dove le caratteristiche di ogni nodo sono indicate da h_i.

E=$\{(e_k, r_k, s_k)\}$ è l'insieme di archi in cui ogni nodo è rappresentato dalle proprie caratteristiche e_k, il nodo che riceve r_k, ed il nodo che invia s_k.

u denota un insieme di attributi globali che rappresentano lo stato del sistema sottostante.

Ogni blocco GN consiste di sei funzioni

Tre di aggiornamento φ

$$\mathbf{e}'_k = \phi^e\left(\mathbf{e}_k, \mathbf{h}_{r_k}, \mathbf{h}_{s_k}, \mathbf{u}\right)$$
$$\mathbf{h}'_i = \phi^h\left(\bar{\mathbf{e}}_k, \mathbf{h}_i, \mathbf{u}\right)$$
$$\mathbf{u}' = \phi^u\left(\mathbf{e}', \mathbf{h}', \mathbf{u}\right)$$

Tre di aggregazione ρ

$$\bar{\mathbf{e}}'_i = \rho^{e\to h}\left(E'_i\right)$$
$$\bar{\mathbf{e}}' = \rho^{e\to u}\left(E'\right)$$
$$\bar{\mathbf{h}}' = \rho^{h\to u}\left(H'\right)$$

E'$_i$=$\{(e'_k, r_k, s_k)\}$ è l'arco con le caratteristiche dei nodi aggiornate

E'=$\cup_i$E'$_i$=$\{(e'_k, r_k, s_k)\}_{k=1:N_e}$ è l'insieme degli archi con le caratteristiche dei nodi aggiornati

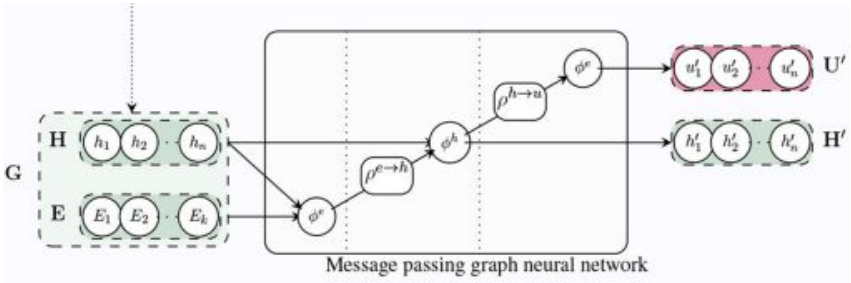
H'=$\{(h'_i)\}_{i=1:N_v}$ è l'insieme dei nodi con caratteristiche aggiornate.

Per effettuare l'addestramento completo di un blocco GN, sono necessari 6 steps, alternando l'aggiornamento con l'aggregazione.

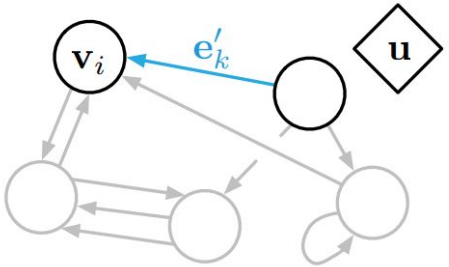
Graph Neural Network [3]

In linea generale quando un grafo G, viene dato in input ad un blocco GN, la computazione procede dal lato, al nodo fino al livello globale.

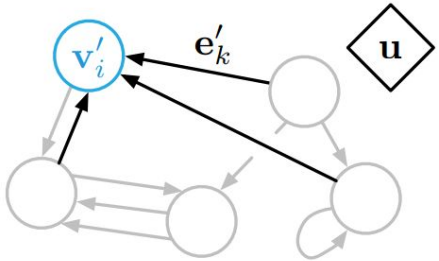
Nelle immagini riportate sulla destra, l'azzurro indica gli elementi che stanno per essere aggiornati, mentre il nero indica gli altri elementi che sono coinvonti nell'aggiornamento



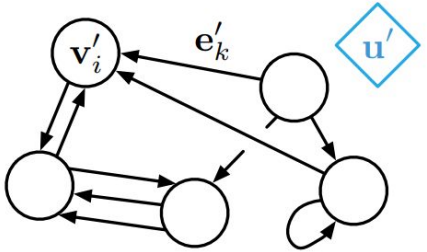
Aggiornamento dell'arco



Aggiornamento del nodo



Aggiornamento globale



Scenario clinico [1]

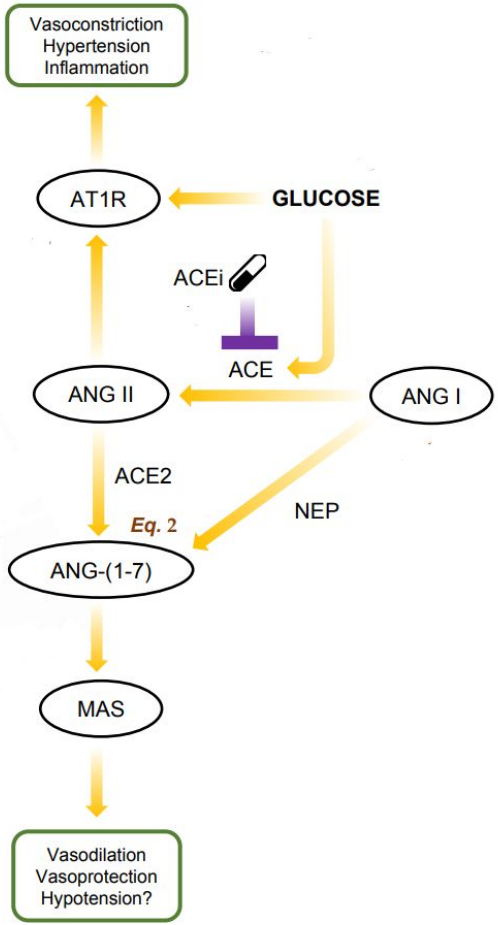
Lo scenario clinico consiste in un paziente anziano che soffre di ipertensione, diabete ed ha contratto un'infezione da SARS-CoV 2.

Scenario clinico [2]

Esiste una relazione tra ipertensione, diabete e COVID? **SI**

Il **sistema renina-angiotensina** (RAS) è un sistema ormonale che regola la vasocostrizione. Il principale regolatore del RAS è l'ormone peptidico **Angiotensina II** (ANG II). L' ANG II esercita le proprie funzioni mediante due recettori AT1R e AT2R, e mediante l'Angiotensina (1-7) (ANG 17), prodotto dall'enzima di conversione ACE2 che attiva il recettore MAS. Questi tre recettori si occupano di regolare la pressione sanguigna facendo cose differenti:

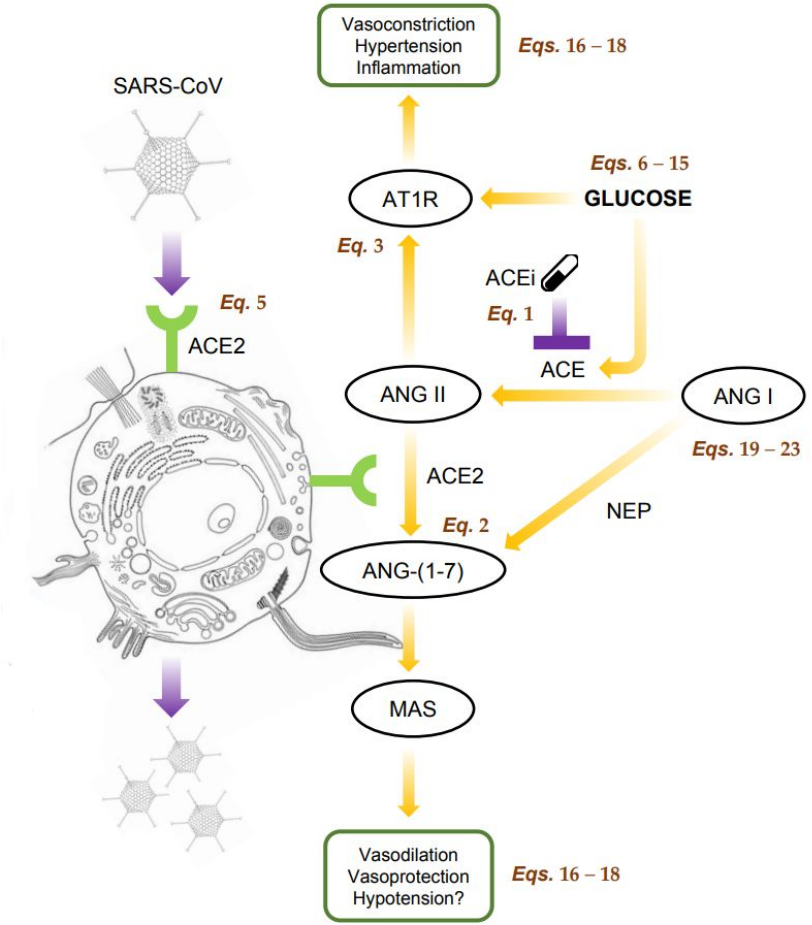
- ATR1: induce la vasocostrizione e dunque l'ipertensione
- ATR2: comportamento controverso, infatti in condizioni fisiologiche normali contrasta gli effetti del AT1R, tuttavia gli effetti vasodilatatori che produce non vengono associati a una significativa riduzione della pressione sanguigna
- MAS: promuove la vasodilatazione e l'abbassamento della pressione sanguigna



Scenario clinico [3]

Il punti critici che legano l'ipertensione a diabete e covid sono:

- Un'elevata concentrazione di glucosio causata dal diabete, può determinare condizioni di ipertensione.
- Un'infezione da COVID-19 può avere ripercussioni sulla pressione sanguigna poiché il virus si lega all'ACE 2 e ne impedisce la corretta attività, ovvero la generazione dell'ANG 17.



Aspetti implementativi [1]

Dataset	Tipo Features	Numero Features	Numero Sample
CARDIO	Sistema Cardiaco	27	1809
RAS	Sistema renina-angiotensina	13	2501

- Il dataset **CARDIO** contiene features i cui valori sono pressioni misurate in millimetri di mercurio, più una feature che rappresenta il tempo.
- Il dataset **RAS** contiene features i cui valori sono concentrazioni misurate in (ng/mL), più una feature che rappresenta il tempo.

Il **primo passo** è la riduzione della dimensionalità per mantenere solo le features utili alla sperimentazione:

Delle 27 features di CARDIO vengono mantenute:

- **t2**: il tempo
- **Pra,Prv,Pla,Plv**: rispettivamente pressione dell'atrio e ventricolo destro e sinistro

Delle 13 features di RAS vengono mantenute:

- **t1**: il tempo
- **angI,Inhibition,Renin,AGT,angII,diacid,ang17,at1r,at2r,ACE2**: valori relativi al sistema renina-angiotensina

Aspetti implementativi [2]

Il **secondo passo** è quello di mettere in relazione il tempo con i valori di pressione (per CARDIO) ed il tempo con i valori di concentrazione (per RAS).

Questa operazione viene fatta andando ad interpolare la colonna del tempo del dataset RAS con tutte le colonne relative al RAS, e la colonna del tempo del dataset CARDIO con tutte le colonne relative al CARDIO.

```
x_list = []
    for c in list(x_ras.columns) + list(x_cardio.columns):
        if c in x_ras.columns:
                f = interpolate.interp1d(tx_ras, x_ras[c].values)
                x_list.append(f(t1))
        elif c in x_cardio.columns:
                f = interpolate.interp1d(tx_cardio, x_cardio[c].values)
                x_list.append(f(t2))
```

Aspetti implementativi [3]

Il **terzo passo** è quello di separare i samples dalle labels:

```
for batch in range(x.shape[0] - 2 * window_size + 1):
    samples.append(x[batch:batch + window_size - 2])
    labels.append(x[batch + window_size - 1:batch + 2 * window_size - 3])
```

Ogni sample così come ogni label sarà una lista di valori:

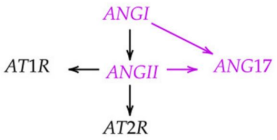
- Il primo sample saranno i valori x[0:998], la prima label saranno i valori x[999:1997]
- L'ultimo sample saranno i valori x[4000:4998], l'ultima label saranno i valori x[4999:5997]

Il **quarto ed ultimo passo** è quello di andare a partizionare i dati, in dati di training e dati di test ed allenare il modello. Il partizionamento veniva fatto mediante un blocco di codice poco leggibile e confusionario, per questo è stato sostituito mediante una funzione di train_test_split. In particolare i dati vengono partizionati i 70% per il training e 30% per il test. Per quanto riguarda l'addestramento sono state inoltre usate 20 epoche ed un learning rate di 0.01.

Aspetti implementativi [5]

Prima parlavamo di un grafo che veniva dato in input ad un blocco GN. Nella figura sulla destra è possibile vedere nel concreto la “elist” cioè la lista di archi dalla quale viene creato il grafo.

La modellazione del grafo viene effettuata sfruttando conoscenze mediche. Infatti i nodi corrispondono alle variabili rappresentate dalle equazioni differenziali presenti nel lavoro di Pietro Barbiero e Pietro Lio (The Computational Patient has Diabetes and a COVID).



$$\frac{d[ANGI]}{dt} = k_{NEP}[ANGI] + k_{ACE2}[ANGII] - \frac{\ln 2}{h_{ANG17}}[ANG17]$$

$$\frac{d[AT1R]}{dt} = (a_{AT1R}G + b_{AT1R})[ANGII] - \frac{\ln 2}{h_{AT1R}}[AT1R]$$

$$\frac{d[AT2R]}{dt} = k_{AT2R}[ANGII] - \frac{\ln 2}{h_{AT2R}}[AT2R]$$

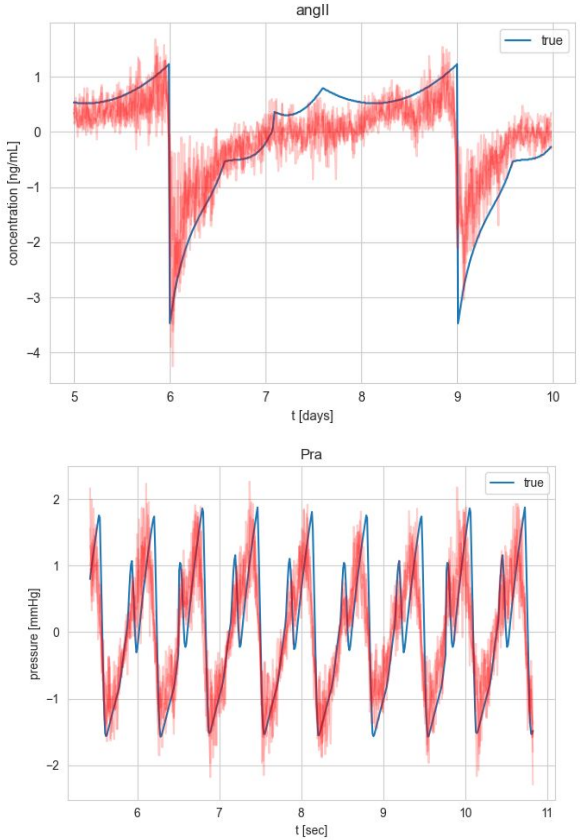
```
elist = [
    #colonne relative a x_ras lega ognuno con se stesso (crea un nodo)
    ('t', 't'), ('angI', 'angI'), ('Inhibition', 'Inhibition'),
    ('Renin', 'Renin'),('AGT', 'AGT'), ('angII', 'angII'),
    ('diacid', 'diacid'), ('ang17', 'ang17'), ('at1r', 'at1r'),
    ('at2r', 'at2r'), ('ACE2', 'ACE2'),

    #colonne relative a x_ras lega il tempo con tutti
    ('t', 'angI'), ('t', 'Inhibition'), ('t', 'Renin'), ('t', 'AGT'),
    ('t', 'angII'), ('t', 'diacid'),('t', 'ang17'), ('t', 'at1r'),
    ('t', 'at2r'), ('t', 'ACE2'),

    #colonne relative a x_ras legami angI
    ('AGT', 'angI'), ('Renin', 'angI'), ('angI', 'ang17'),
    ('angI', 'angII'),

    #colonne relative a x_ras legami angII
    ('diacid', 'angII'),('angII', 'Renin'),('angII', 'ang17'),
    ('angII', 'at1r'),('angII', 'at2r'),

    #colonne relative a x_ras legami ACE2
    ('ACE2', 'ang17'), ('ACE2', 'angI'),

    #colonne relative a x_cardio lega tempo con tutti
    ('t2', 'Pra'), ('t2', 'Prv'), ('t2', 'Pla'), ('t2', 'Plv'),

    #colonne relative a x_cardio lega tutti 4 valori restanti fra loro
    ('Pra', 'Prv'), ('Pra', 'Pla'), ('Pra', 'Plv'),
    ('Prv', 'Pra'), ('Prv', 'Pla'), ('Pra', 'Plv'),
    ('Pla', 'Pra'), ('Pla', 'Prv'), ('Pla', 'Plv'),
    ('Plv', 'Pra'), ('Plv', 'Prv'), ('Plv', 'Pla'),
        ]
```

Risultati

Una volta allenato il modello viene utilizzato, come visibile sulla destra, per generare un fascio di possibili traiettorie per gli elementi del set di test.

L'accuratezza del modello realizzato da Barbiero et al., cioè l'accuratezza delle possibili traiettorie è del 95% per ogni variabile (angII, angI, ang17, Pra, Prv, etc).

In particolare quelli visibili sulla destra sono i risultati relativi all'Angiotensina II e alla Pressione dell'Atrio Destro.



3

Bioinformatica sul Cloud

1. Perché il Cloud: il caso di AWS
2. Containerizzare un progetto di Bioinformatica con Docker
3. Creare una routine di CI/CD per un progetto Python di ML
4. Interagire con il modello di ML in Flask
5. Sviluppi futuri

Il Cloud e la Bioinformatica (1)

- **La bioinformatica presenta *tante* sfide.**
- **Il Machine Learning**, come abbiamo visto, è uno dei **campi applicativi che meglio riescono a realizzare soluzioni** eleganti ai **problemi moderni della bioinformatica**.
- Tuttavia, nessuno è un "*one-man-army*". La **progettazione ed implementazione di modelli di machine learning** è un'attività **condivisa, partecipata, collaborativa**.

Cosa significa?

Sviluppare strumenti di Machine Learning nel mondo enterprise implica, molto spesso:

- **Terabyte** (se non esponenzialmente di più) **di dati**
- **Grandi team** di scienziati ed ingegneri
- **Codebase** articolate, complesse, **multi-repository**
- **Architetture distribuite o a microservizi**.
- Difficile ed **oneroso accesso alle dipendenze di sviluppo**
- Costi notevoli per servire l'utente finale

... ma soprattutto ...

- **Difficile deploy e fruizione.** Due problemi:
  - Come fa **il team** a lavorare al modello di machine learning in maniera *seamless*?
  - Come fa **l'utente finale** ad interagire con il modello di machine learning?

Il Cloud e la Bioinformatica (2)

Il team

- In generale, quando un team implementa ***insieme*** un modello di machine learning e i vari tool ad esso associati, nascono vari problemi.

- L'*i*-esimo membro del team, infatti, deve:

    - Scaricare sul suo computer una **quantità esorbitante di dati** (spesso migliaia di file).
      E se non avesse memoria disponibile?
      E se si trattasse di decine di terabyte di dati?

    - Assicurarsi di installare tutte le corrette **dipendenze di Python** (es. numpy, pandas) …
      E se avessi più versioni di Python o lavorassi a più progetti di machine learning con necessità distinte?

    - Assicurarsi di avere il **sistema operativo** giusto.
      Se il team opera su Linux, ma io uso Windows… che si fa? ***Virtualizzare*** *non è un'opzione. Troppo dispendioso in termini di risorse.*

    - Dopo l'implementazione, fare versioning ed **impiegare risorse personali per runnare il modello e garantire l'integrità** del progetto.
      E se pushassi su GitHub codice malevolo, corrotto, non funzionante? Potrei rischiare di auto-sabotare l'intera rete neurale.

[!] Tutto ciò è **lento**, **inefficiente**, **costoso** (il personale si paga, soprattutto quello iper-specializzato che si occupa di ML applicato alla bioinformatica).

Non va bene. **Il team**, inteso come "risorsa umana", **dev'essere scorporato dai requisiti lavorativi**.

Il Cloud e la Bioinformatica (3)

Problemi e soluzioni

- Non vogliamo che i membri del team debbano:
  - **scaricare enormi quantità di dati**
  - **"contaminare" il proprio computer con molteplici versioni differenti dei tool di sviluppo**
  - **costringerli a cambiare sistema operativo** o - peggio ancora - virtualizzare un altro OS e installare sopra le dipendenze, operazione che ha requisiti hardware non trascurabili
  - **si sentano responsabile per codice difettoso** quando, in realtà, è il sistema a doversi occupare di verificarne l'integrità
  - **debbano impiegare le proprie risorse di calcolo** per effettuare training, testing, etc

In particolare, la problematica (5) ben si coniuga con un'altra necessità:

SCALARE VERTICALMENTE

Fare Machine Learning richiede molta potenza di calcolo.

Non è detto che il team stia lavorando su computer adeguatamente performanti. E non è neanche detto che l'azienda abbia a disposizione centinaia di migliaia di euro da spendere in attrezzatura per i dipendenti.

Si vuole trovare una **soluzione che risolva i problemi lavorativi del team**, e che permetta al **modello di machine learning** di **usare tutta la potenza di calcolo necessaria**...

... *la soluzione è intuibile* ...

Vantaggi e possibili provider

Il Cloud: una overview

Sappiamo cos'è. Ma di preciso **cosa vogliamo dal Cloud**?

- **Computazione elastica.**
- **Opzioni di storaging flessibili.**
- **Costi contenuti e in base alla necessità.**
- **Piattaforme di containerizzazione.**
- **Metodi di deploy rapidi.**

... e inoltre, anche se non è scontato ...

- [Sviluppi futuri] Vulnerability Assessment / Auditing
- [Sviluppi futuri] Orchestrazione intercontinentale
- [Sviluppi futuri] Versioning dell'intera infrastruttura

Google Cloud Platform

IBM Cloud

Computazione Elastica
Infrastructure as a Service

- *Computazione elastica.*
- **Opzioni di storaging flessibili.**
- **Costi contenuti e in base alla necessità.**
- **Piattaforme di containerizzazione.**
- **Metodi di deploy rapidi e robusti.**

Bioinformatica sul Cloud
Storage as a Service

- *Computazione elastica.*
- *Opzioni di storaging flessibili.*
- **Costi contenuti e in base alla necessità.**
- **Piattaforme di containerizzazione.**
- **Metodi di deploy rapidi e robusti.**

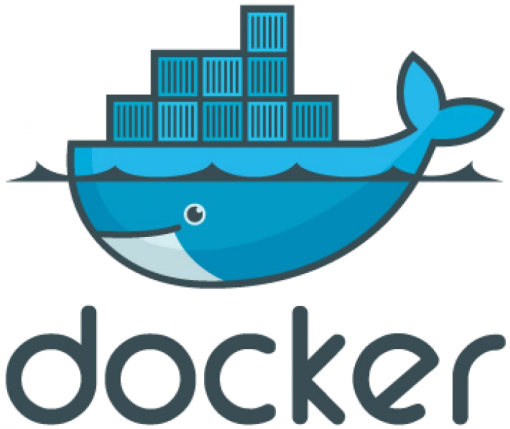
Bioinformatica sul Cloud
Serverless Bioinformatics

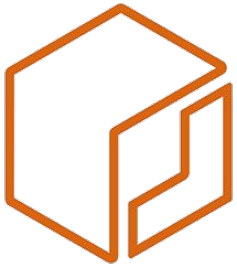
- *Computazione elastica.*
- *Opzioni di storaging flessibili.*
- *Costi contenuti e in base alla necessità.*
- **Piattaforme di containerizzazione.**
- **Metodi di deploy rapidi e robusti.**

AWS Lambda

Bioinformatica sul Cloud
Containerizzazione

- *Computazione elastica.*
- *Opzioni di storaging flessibili.*
- *Costi contenuti e in base alla necessità.*
- *Piattaforme di containerizzazione.*
- **Metodi di deploy rapidi e robusti.**

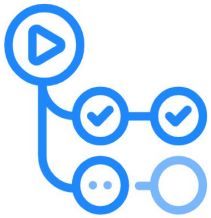




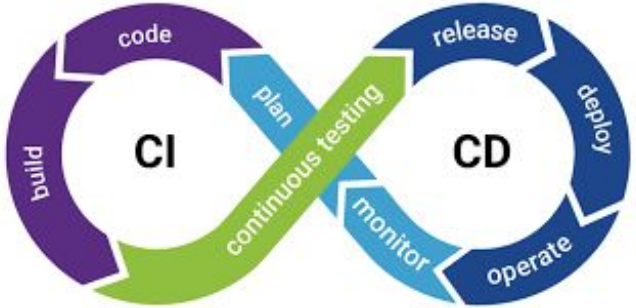
Amazon ECR

Bioinformatica sul Cloud
Deploy e routine di CI/CD

- *Computazione elastica.*
- *Opzioni di storaging flessibili.*
- *Costi contenuti e in base alla necessità.*
- *Piattaforme di containerizzazione.*
- *Metodi di deploy rapidi e robusti.*



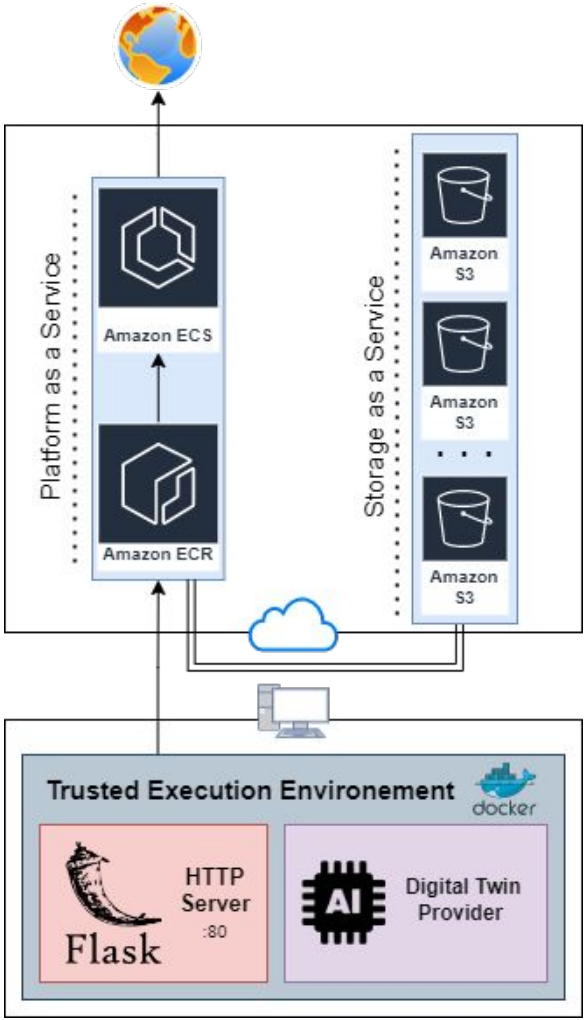
GitHub Actions



La nostra architettura cloud

L'architettura in sintesi

- **Immagine Docker su AWS ECR**
  - **Provider di un Digital Twin** "switchable"
  - **Webapp in Flask** per offrire un minimo grado di interazione con il modello di ML
- **Container Docker su AWS ECS**
- **Orchestrazione Serverless**/FaaS **via AWS Fargate**
- **Dataset su AWS S3** (Storage as a Service)
- **Routine di CI/CD** "on push" su GitHub Actions



Demo

Basta chiacchiere.

E’ tempo di una demo!

Overview infrastruttura AWS, repository GitHub, webapp

Sviluppi futuri Kubernetes & Serverless Framework

- **Orchestrazione multi-cloud**
- **Scalabilità orizzontale**
- **Infrastructure as Code (DevOps)**

Grazie a tutti per l’attenzione!

github.com/antoniogrv/mlops-patient-digital-twin