PLANET: **P**ollution **L**ake **AN**alysis for **E**ffective **T**herapy

Elisabetta Ronchieri (slide di Daniele Spiga)
INFN-CNAF 13.7.2020

Outline

- Motivazioni : **Perchè ?**
- La strategia : **Come ?**
- Workplan e Timeline in sintesi: **Quando ?**
- Personale e sezioni: **Chi ?**
- Il budget: **Quanto costa ?**

- Alcune riflessioni
- Il progetto PLANET e la call FISR

Motivazioni (Perchè?)

- La pandemia COVID-19 è un grave problema di salute globale
- Diffusione estremamente eterogenea su territorio nazionale
  - maggiore frequenza e gravità di casi in regioni o zone a maggior tasso di industrializzazione/traffico veicolare
    - avvalora l'ipotesi che l'inquinamento atmosferico contribuisca alla diffusione e alla severità
- 2 meccanismi potenzialmente sinergici:
  - **Cronico**: l'esposizione cronica agli inquinanti potrebbe indurre danno polmonare, esacerbando le conseguenze dell'infezione virale
  - **Acuto**: il microparticolato derivato dai combustibili fossili, fungendo da "trasportatore", aumenterebbe percorrenza e permanenza in aria del virus, aumentando le possibilità di diffusione.

Covid-19 e Inquinamento atmosferico

- Tuttavia la letteratura mostra che il problema relativo al ruolo dell'inquinamento nell’infezione Covid-19 è ancora in fase di studio
  - gli esperti contestano persino la possibilità di diffusione del bioaerosol del virus SARS-CoV-2

- Risultati di studi inclusivi di quegli elementi che possano limitare l’effetto dei confondenti non sono ancora pubblicati

- Il “lockdown” ha rappresentato, involontariamente, l’esperimento su scala più grande di sempre di riduzione delle emissioni seppure in maniera disomogenea rispetto al tipo di inquinante
  - Potrebbe rappresentare una opportunità unica per il perfezionamento e la validazione delle ipotesi fatte

Quindi: Il progetto PLANET

- PLANET propone di effettuare un’analisi su un insieme molto ampio di dati eterogenei
  - ambientali (relativi all’inquinamento e al clima)
  - clinici (inclusivi di elementi ritenuti determinanti per l'incidenza e la mortalità del Covid-19)
  - Propone inoltre di introdurre livelli di dettaglio individuale o di microaree
  - nonchè di misuratori finalizzati ad arricchire i dataset, come ad esempio la comorbosità e le condizioni socio-economiche/industriali.

Aspetto caratterizzante:

- Si vuole limitare il più possibile l’effetto dei confondenti (correlazioni non-sense)
  - La varietà dei dati, e la granularità (microaree) aiuterebbe, non solo e non primariamente, a valutare l’effetto di causalità ma è ritenuto un elemento chiave per la rivelazione (e predizione) di suscettibilità e vulnerabilità

- Elemento essenziale per analisi predittive

Strategia (come?)

La metodologia adottata da PLANET si basa sulla centralità del dato, inteso come una molteplicità di informazioni eterogenee, di natura diversa.

Oltre alla rete "stazione atmosferica" ARPA e al registro nazionale COVID, alcune delle sorgenti dei dati ritenute accessibili anche grazie alla collaborazione diretta dell'ospedale di Perugia, sono:

- Registro nazionale dei certificati di morte;
- Registri regionali dei ricoveri ospedalieri, che riportano anche dettagli quali la codifica delle malattie
- Informazioni provenienti dal sistema di Geocodifica degli assistiti (disponibile per la regione umbria ma esportato ed esportabile ad altre regioni)
- Database nazionale delle infezioni del virus respiratorio INFLUNET

Ma anche:

- caratteristiche meteorologiche, il contesto urbano verso quello rurale, la densità della popolazione, la densità degli impianti industriali e loro caratteristiche principali, nonché la morfologia del territorio
- aspetti socio-economic/sociali

Strategia … (cont)

Per quanto riguarda la gestione dei dati, la loro aggregazione (ingestion) e organizzazione all'interno di un sistema di tipo DataLake, si prevede di procedere con una strategia basata sul modello dello **"schema on read"**(*) combinato alla creazione "automatica" del catalogo (i.e. **gestione metadati**)

Perchè:

- permette di includere dati eterogenei (tipicamente non strutturati o semi strutturati)
- non richiede la comprensione anticipata dei dati e consente di soddisfare i requisiti futuri, che in questa tipologia di analisi/studi non sono totalmente prevedibili
- permette di implementare un sistema generico e multidisciplinare, da mettere a fattore comune in forma di Toolkit multidisciplinare su infrastruttura INFN-Cloud (con significativo impatto a livello nazionale).

(*) rispetto allo "schema on write" permette l'applicazione dello schema in fase di lettura, in funzione dell'esigenza del momento, permettendo di creare nuovi schemi, quando necessario, da applicare alla fase successiva di data modelling

Obiettivi principali

- Aumentare la conoscenza dei meccanismi di diffusione del virus SARS-CoV-2, fornendo importanti dati per la modellistica
- Fornire supporto diretto all'ipotesi, finora induttiva, secondo la quale è necessario ridurre l'uso di combustibili fossili, per migliorare la salute pubblica
- Individuare i dati in grado di migliorare i processi decisionali clinici e di politica sanitaria e ambientale.
- Sviluppare una piattaforma multidisciplinare per il trattamento dei dati eterogenei

Caratteristiche di PLANET

- Studio a variabili multiple finalizzato alla valutazione dell'incidenza dell'inquinamento atmosferico sull'infezione da COVID-19 e alla sua gravità
  - **Aggregazione di molteplici fonti informative**
- **Gestione di tipologie di dati eterogenei,** strutturati, semi strutturati e non strutturati
- **Utilizzo della infrastruttura di calcolo INFN-Cloud** e di quella certificata ISO/IEC 27001 del CNAF
- **Estensione di servizi software di alto livello derivati dall'esperienza INFN nel calcolo distribuito**
  - sistema di storage, aggregazione e conservazione dei dati processamento dei dati
- **Riuso di soluzioni sviluppate e fornite da progetti di ricerca dell'INFN**,
  - come INDIGO-DataCloud, EOSC-hub, XDC, IoTwins, ML-INFN, Harmony
- **Fruizione delle competenze dei partecipanti:**
  - nel campo dell'epidemiologia, della sanità personalizzata
  - gestione e sviluppo di infrastrutture distribuite, gestione e analisi dati, integrazione e sviluppo di sistemi di calcolo

Workplan (in sintesi)

Il progetto è supposto avere una durata di due anni, tempo ritenuto necessario dai proponenti per permettere il completamento delle attività descritte.

In particolare la strategia identificata per lo sviluppo di PLANET si articola in due fasi:

1. si concentrerà sulla creazione di un prototipo completo su scala regionale integrato con il sistema di geocodifica, già disponibile in Umbria per lo studio di microaree, che utilizzerà dati legati a informazioni cliniche, ambientali e ai misuratori descritti.
2. sarà finalizzata all'estensione del pilota multivariabile ad altre regioni, incluso del supporto di geocodifica, rendendo possibile l'analisi completa e integrata su ampia statistica e quindi creazione della struttura finale di un "Data Lake" nazionale.

Timeline (quando?)

Primo Anno

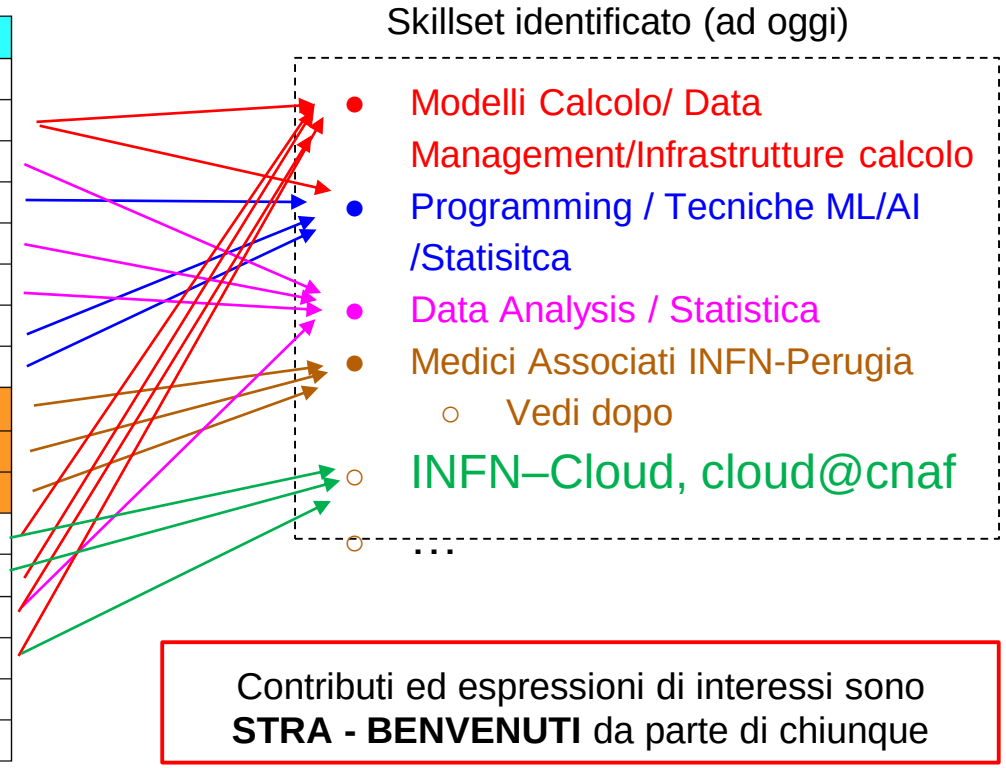
1. PM 4: Analisi dei requisiti
2.
3. PM 6: implementazione completa dei servizi di gestione di metadati integrata nel prototipo
4. PM 6: Prima analisi su dati dettagliati a scala regionale (Umbria)
5. PM 12: Analisi a scala nazionale (regioni selezionate) su dati parziali e finalizzazione della raccolta dati di dettaglio
6. PM12: Risultati di test di scala del datalake comprensivo del sistema di gestione metadati

Secondo Anno

1. PM18: Implementazione e integrazione con automatismi Cloud (Templates e setup automatizzati per replicabilità estendibilità)
2. PM18: Automazione dei processi di gestione di metadati (creazione catalogo) e automazione dei processi di gestione data cleaning/validation
3. PM 24: Validazione tecnica del sistema integrato
4. PM 24 Risultati analisi completa su scala nazionale Lombardia, Veneto, Liguria, Toscana, Umbria.
5. PM 24 Documentazione completa

Personale Coinvolto (Chi ?)

Sede	Personale	Percentuale afferenza
PERUGIA	Daniele Spiga	25%
	Diego Ciangottini	10%
	Pasquale Lubrano	10%
	Loriano Storchi	10%
	Matteo Duranti	0%
	Sara Cutini	0%
	Tommaso Tedeschi	10%
	Mirco Tracolli	10%
	Giuseppe Ambrosio	15%
	Giampaolo Reboldi	15%
	Fabrizio Stracci	15%
CNAF	Davide Salomoni	5%
	Cristina Duma	10%
	Elisabetta Ronchieri	10%
	Alessandro Costantini	10%
TOTALE FTE		1.45



Contributi ed espressioni di interessi sono **STRA - BENVENUTI** da parte di chiunque

Componente Medica: “bio in pillole”

Prof. Ambrosio:

- Direttore del maggior Dipartimento ad impronta biomedica, all’interno del quale si ritrovano ulteriori competenze utilizzabili per ulteriori approfondimenti e valutazioni dei risultati del presente progetto (es: Microbiologia/Virologia, Malattie Infettive, Medicina di Urgenza, etc…).
- Già Direttore Sanitario, e Direttore Generale ad interim, dell’Azienda Ospedaliera di Perugia. Apportando così competenze specifiche di Sanità Pubblica, e management dei sistemi sanitari.
- Componente del Comitato Tecnico-Sanitario del Ministero della Salute.
- Componente della Task Force China-Europa per lo studio dell’impatto del COVID su pazienti cardiopatici.

Prof. Fabrizio Stracci:

- Associato di Igiene. Esperto di Sanità Pubblica e modelli epidemiologici. Responsabile del Registro Regionale pazienti oncologici, che traccia e monitora questi pazienti su tutto il territorio regionale. Componente del Comitato Tecnico-Scientifico COVID della Regione Umbria.

Prof. GianPaolo Reboldi.

- Associato di Nefrologia, Esperto di Metodologia della Ricerca clinica e Farmaco-Epidemiologia. Componente del Comitato Registri della Società Italiana di Nefrologia.

Il Budget (quanto costa ?)

Il progetto non prevede **nessun budget per materiale inventariabile**

- il primo anno si intende implementare il prototipo usando le infrastrutture esistenti
  - in particolare **INFN Cloud e cloud@cnaf** su scala ridotta per mettere a punto tutte le funzionalità del servizio e dei modelli di analisi dei dati.
- Si **richiedono 2k Euro di missioni per sezione** al fine di consentire qualche missione CNAF <-> PG e la partecipazione ad una conferenza/workshop
  - Infrastrutturale/data science e/o di dominio

Per il **secondo anno si prevede la possibilità di un piccolo finanziamento per potenziare il sistema di storage**

- O(10K Euro) che saranno gestiti al CNAF.

Richieste e impatto sulla sezione

Non è prevista nessuna richiesta sui servizi della sezione

- La parte infrastrutturale sarà collocata su risorse di

  Cloud@CNAF e INFN-Cloud (*)

(*) ne abbiamo parlato o ne parleremo in altro talk durante questo CdS

Alcune riflessioni: Potenziali impatti

- FISR Sara Cutini
  - ML/DL analysis su dati clinici finalizzato alla comprensione dell'infezione Covid-19
    - Se finanziato potrebbe usufruire della piattaforma di data science di PLANET
- ERC Advanced Prof. Luca Castelli (Professore associato di Diritto pubblico presso Dipartimento di Economia dell'Università di Perugia)
  - INFN (PG + CNAF) partecipa come terza parte (third party: 120k personale + 40k HW @cnaf) per:
    - sviluppo e provisioning di piattaforma (HW+SW) per il trattamento di dati eterogenei integrato su INFN-Cloud
    - Sviluppo modello di calcolo
    - Data processing
  - [Altra third party è USL (Dott. Carla Bietta)]

→ Se finanziato potrebbe usufruire della piattaforma di data science PLANET

Perchè una sigla in CNS5 ?

- E’ una attività multidisciplinare
  - Nasce dalla richiesta di supporto da parte di Medici di UniPG
- Ha una forte componente di infrastruttura/ servizi di calcolo
  - La CCR non finanzia questo tipo di progetti
- L’attività proposta presenta forti sinergie con attività INFN (anche finanziate da CNS5)
  - ML_INFN, AIM !?!
  - E non solo: INFN-Cloud …
  - Potenzialmente il servizio di piattaforma Data Lake realizzato da PLANET può diventare un altro componente del toolkit INFN messo a fattore comune
- Può evolvere verso un sistema a supporto di altre analisi metodologicamente analoghe
  - O anche fungere da piattaforma base per l’implementazione di sistemi predittivi

PLANET come progetto FISR

A Giugno abbiamo deciso di sottomettere un progetto alla call FISR:

"Il MUR intende acquisire e selezionare proposte progettuali di ricerca di particolare rilevanza strategica, finalizzate ad affrontare le nuove esigenze e questioni sollevate dalla diffusione del virus SARS Cov 2 e della infezione Covid 19 - Decreto Direttoriale n.562 Covid 2020"

Titolo Progetto:
Studio e previsione di effetti acuti e cronici dell'inquinamento atmosferico sul COVID-19 in Italia con analisi di dati provenienti da sorgenti eterogenee in un Data Lake federato e certificato ISO.
Partners: INFN e UniPG (dipartimento di medicina)
Budget: 80 MUR + 20% cofinanziamento

- 54 K INFN
- 14 K UniPG
- 15 K Consulenze (ICT4life srl / IRCCS MultiMedica Spa)

Durata: 6 mesi (+ 6 mesi)
PI: INFN (Daniele Spiga)
Sezioni coinvolte INFN: Perugia e CNAF