# PLANET ( preventivi 2022 )

Slides courtesy of Daniele Spiga
(INFN-PG, Resp. Nazionale)

# Cosa è PLANET ( quick reminder )

## L'obiettivo ( ambizione ) del progetto PLANET

Effettuare uno studio di tipo osservazionale ( ecologico ) e valutare l'associazione tra Covid-19 e l'esposizione all'inquinamento atmosferico per contribuire alla comprensione del suo ruolo nella **suscettibilità al contagio** e nella **gravità dei sintomi**.

Come:

- **Includendo molteplici fonti di dati:**
  - si ipotizza che dati atmosferici, densità di popolazione, ambiente urbano o rurale, mobilità, condizioni socio-economiche etc influenzino sia i tassi di diffusione e infezione di SARS-COV-2 che la severità. Complementando anche con variabili cliniche ( comorbidità, fragilità.. )
- **Analizzando dati su scala spaziale comunale,** al livello Nazionale

**Obiettivo scientifico**

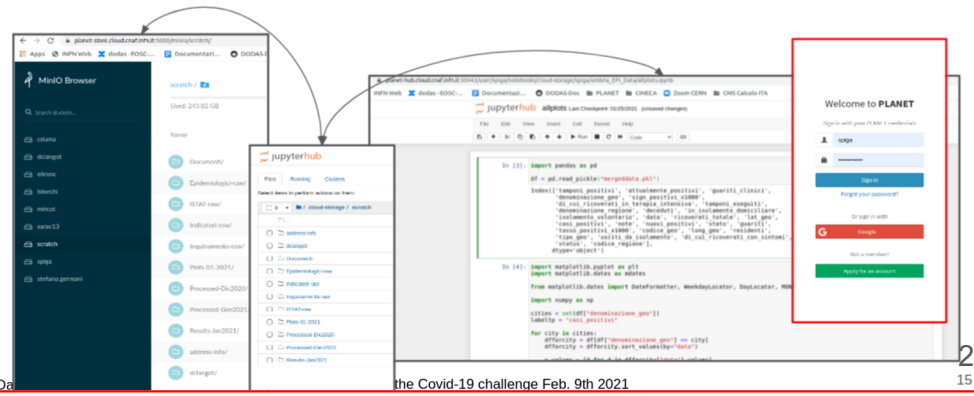
## Obiettivo tecnologico

NOTA BENE: questo si lega fortemente alla problematica ( del t**rattamento dei dati "sensibili"** )

## PLANET **DataLake** (prototipo)

**Integrare e rendere disponibile** una piattaforma "open" e generica (riutilizzabile) per

- Integrazione di dati da sorgenti informative multiple, Processamento, Federazione di risorse attraverso soluzioni di **INFN-Cloud**



# Le sfide poste da questa attività

**La realizzazione dello studio descritto pone molteplici sfide, di diversa natura e con diverso fattore di rischio**

## Raccolta dei dati

**Molteplicità di interlocutori**
- Non esiste referente unico
- **E tantomeno interfacce tecnologiche...**

**Aspetti organizzativi**
- Accordi formali tra enti (DPO), rispetto delle norme, classificazione dei dati, GDPR risk based approach

## Analisi e interpretazione

**Qualità del dato**
- Dati eterogenei e raccolti con finalità diverse dall'uso fatto "Validazione", armonizzazione

**Necessità di Competenze Multidisciplinari**
- Tema ampio che richiede competenze di dominio, bisogna porsi le giuste domande

## Gestione dei dati e delle pipelines

**Infrastruttura (Datalake) scalabile, agile, collaborativa e integrata**
- per la gestione di dati eterogenei

**Procedure specifiche di gestione dei dati sensibili**

Soluzione estendibile
- e riusabile in altri ambiti

# Milestones per il primo anno ( di due )

Le **tre** milestones previste per il 12.2021 sono queste

- **Raccolta dei dati primari a scala nazionale**. Raccolta dei confondenti, in particolare dati clinici per la regione Umbria. I dati saranno integrati nel Data Lake di PLANET.
- **Sviluppo del modello di analisi**, inclusivo dei confondenti estratti dai dati raccolti nella regione Umbria.
- **Prototipo della piattaforma di calcolo** integrante i seguenti componenti: sistema di accesso autenticato , piattaforma di calcolo, sistema di storage. Il sistema ~~di~~ abiliterà l'accesso  ai dati nel formato originale per l'analisi degli stessi. Sarà messa a disposizione una demo dell'intero workflow ovvero delle istruzioni dettagliate per utente finale.

# Dove siamo dopo I primi sei mesi di attività

## Raccolta dei Dati

- Umbria: inquinamento atmosferico “di dettaglio”; Epidemiologici comunali; Indici deprivazione
- Nazionale:
  - Inquinamento Atmosferico ( da Copernicus ); Climatici ( Copernicus: Work in progress );
  - Dati Epidemiologici ISS ( convenzione INFN-ISS ) ; Dati Epidemiologici ASL Viterbo ( convenzione CNAF - ASL ); Dati Epidemiologici Liguria..
  - Indici istat e deprivazione
- **NOTA** il livello di difficoltà sul recupero di dati copre tutto lo spettro: da facile a impossibile…

## Analisi dei dati

- Effettuata validazione dei dati ( primi risultati presentati anche a workshop Covid19 dell’INFN )
- Sono state intraprese/impostate due strade [**vedi dopo**]
  - Studio integrato per l’identificazione delle features ( varibili ) più determinanti; Analisi temporale

## Prototipo della Piattaforma “DataLake”

- Finalizzato il prototipo iniziale
- Definizione e prototipizzazione del sistema finale: Work in Porgress
  - In parallelo qui c’è l’attività di porting in **EPIC** [ vedi dopo ]



Daniele.Spiga@pg.infn.it

# Preparazione dei dati

Oltre il recupero dei dati, abbiamo speso una certa quantità di effort nell'organizzare e validare e comprendere cosa stavamo guardiamo

- Studi degli andamenti degli inquinanti, classificazione per zone geografiche ( i.e per l'umbria collina, valle e conca ), correlazioni tra chimici e particolato, andamenti rispetto agli stati della pandemia etc etc
  - Adesso stiamo validando ARPA vs Copernicus, non è un task banale!

- Senza questo lavoro preliminare tutto il resto non è pensabile

# Analisi dei dati VERY PRELIMINARY

A regression tree basato su un partizionamento ricorsivo in un con
framework. Lo usiamo al fine di ordinare le features ( varia
risposta in base alla loro importanza.

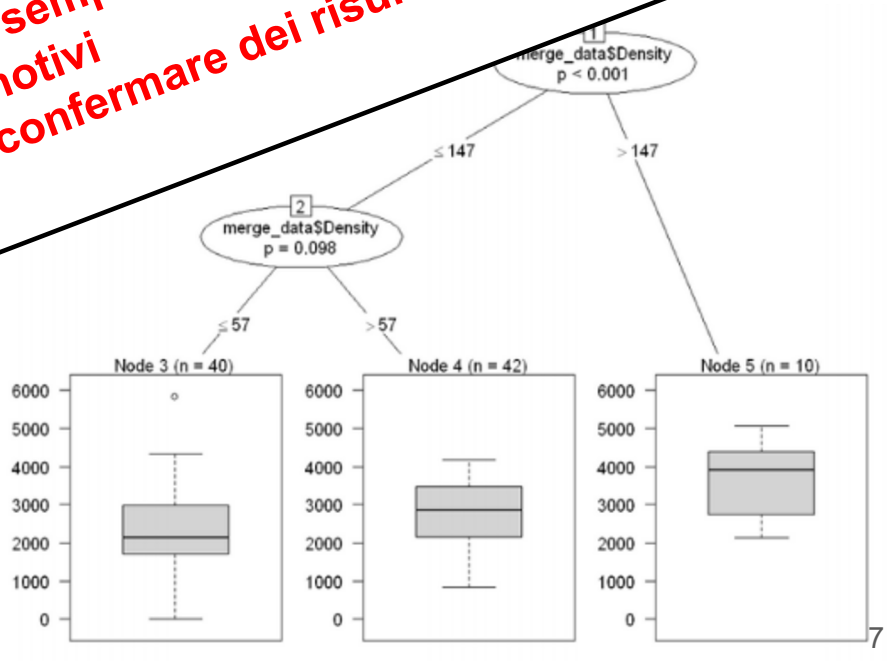
- Questa parte dello studio prende spunto d

**Dataset usato: 92 comuni d**
**e 313 variables:**

- Epidemiol
  ho
- Der
  dens
  social , age
  ranges, ication.
- Pollutan : mean PM10, mean PM2.5, mean NO2, and mean O3

**VERY PRELIMiNARY** Costruendo alcuni modelli ctree e anche clusterizzando (creando sub-groups): **la feature più importante è sempre la densità**

- **Questo non ci sconvolge affatto per vari motivi**
- **Semplicemente al momento sembra non confermare dei risultati in letteratura!**



# Analisi dei dati VERY PRELIMINARY

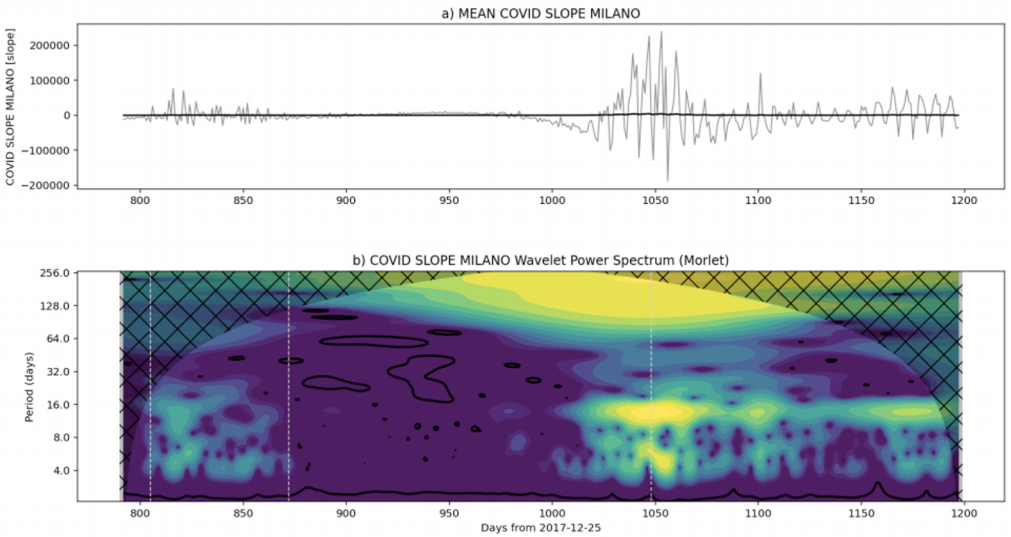
## Abbiamo appena avviato una analisi temporale

- Anche se è solo all'inizio è una delle cose su cui vorremmo investire nei prossimi mesi, anche dati i feedback e le indicazioni degli epidemiologi ( Fabrizio et al. )

Dataset:

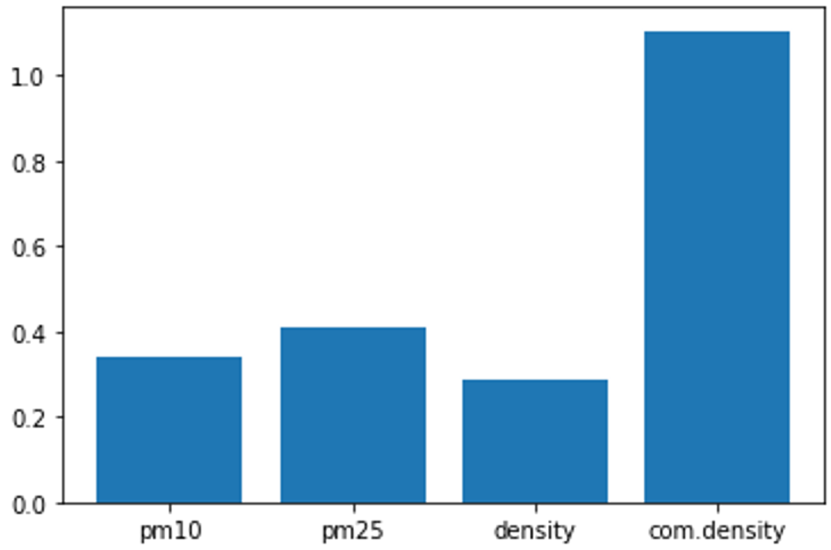
- Copernicus 2017 - 2020
- Dati Covid19: github ma da migrare su ISS [vedi dopo]

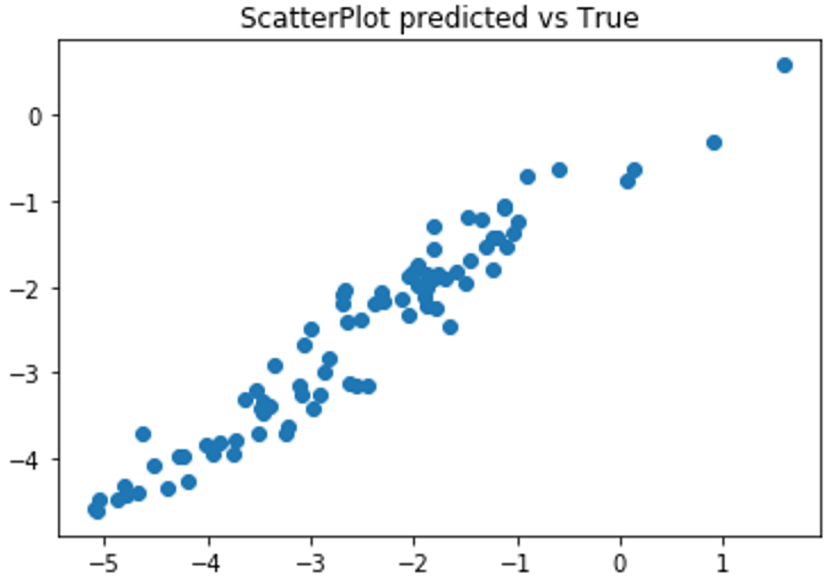
Wavelet Trasform



# Analisi dei dati VERY PRELIMINARY

RandomForest e **Permutation feature importance** per valutare l'importanza delle features




**Dataset**:
**Copernicus** 2017 - 2020

**Dati Covid19**: da Setti L, ey.al. *BMJ Open* 2020;**10:**e039338



Daniele.Spiga@pg.infn.it

# **Analisi dei dati**: cosa c'è nella nostra to do list

Ripetere al livello Nazionale lo studio umbro ( regression tree )

- Includere altri fattori ( commuting … )
- Includere i dati ISS
  - Estendere con control sample come Lazio e Liguria ( sperabilmente altro ?!? )
- Ripetere anche clusterizzazioni per studiare zone omogenee
  - Identificando outliers

Sviluppare lo studio temporale identificando tutti i parametri utili

- Esempio: Ricoveri vs Incidenza vs terapie intensive, mediare sull'età
- Vorremmo considerare anche l'RT al livello provinciale ( possibile sinergia con gruppo CovidStat )

In entrambi i casi va considerata l'inclusione dei parametri climatici

Altre idee… work in progress



Daniele.Spiga@pg.infn.it

# La piattaforma ( DataLake )

**Dove siamo:** La prima versione del prototipo è disponibile ( ed utilizzabile ) qui:

- Data Storage; Data Compute; Autenticazione e autorizzazione dedicata ( ma federata ) .

**Cosa stiamo impostando per la seconda [illegible] stiamo andando):**

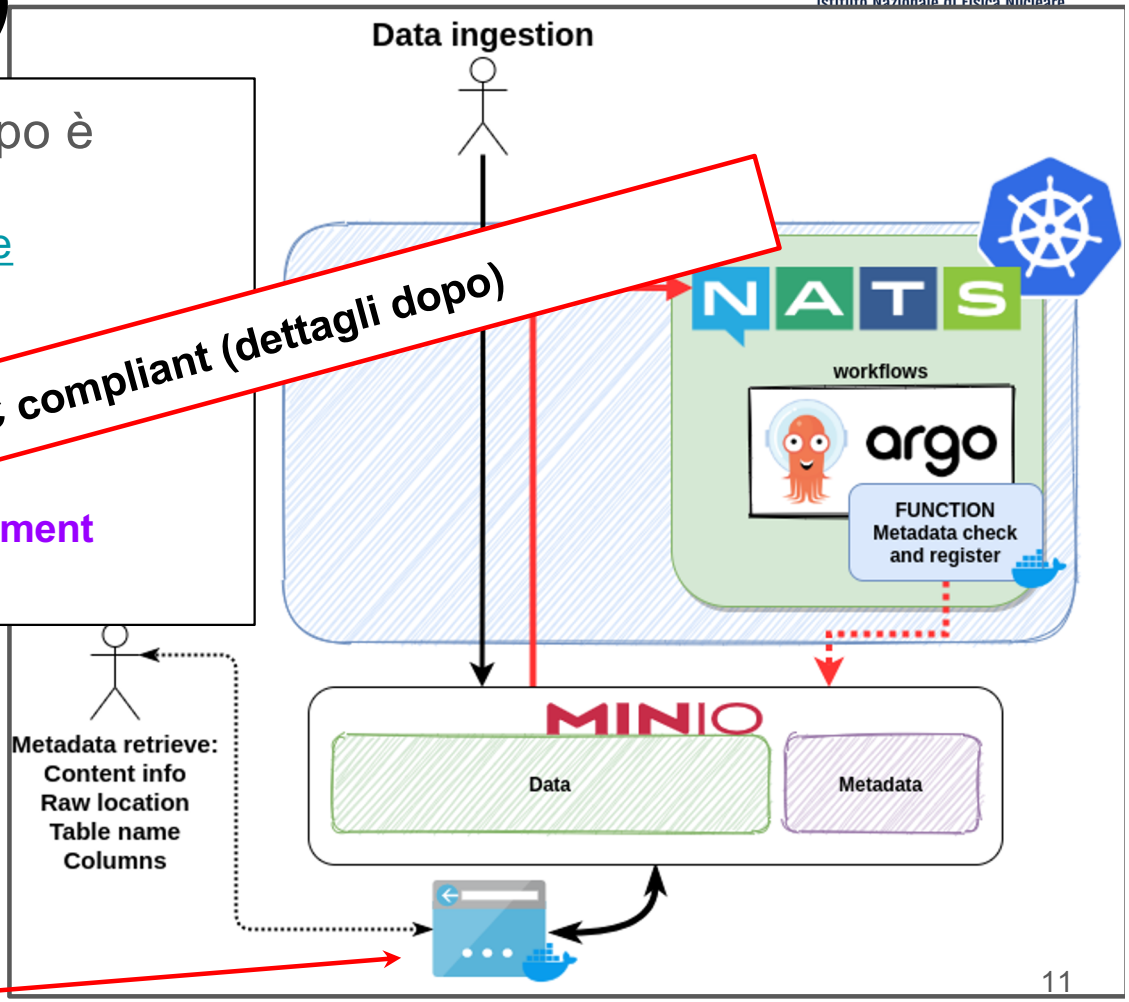
1. Data Ingestion:
2. Quality/[illegible]tadata management
3. C[illegible]i dati

TUTTA la Piattaforma sarà resa EPIC compliant (dettagli dopo)

MinIO

Nats

Argo event + argo workflow




Demo

Click here



Daniele.Spiga@pg.infn.it

/2021

# Cosa è EPIC e perchè EPIC in PLANET

EPIC e' la versione sicura e certificata ISO/IEC 27001 27017 27018 della INFN Cloud

Vantaggi:

- Certezza che tutte le misure tecnico/organizzative previste dal GDPR e dal Garante Privacy siano applicate (i ricercatori che utilizzano dati dell'ISS ed in generale dati personali hanno la responsabilita' di trattare i dati applicando tutte le misure richieste dalla legge)
  - Il sistema viene sottoposto annualmente ad audit da parte di un organismo indipendente
- Minimizzazione dei rischi di data breach attraverso l'adozione di best practice accettate e validate a livello internazionale
- Accesso a servizi e supporto su tematiche legate alla sicurezza delle informazioni: cybersecurity, backup, documentazione per la gestione della confidenzialità' e privacy dei dati, interfaccia verso il DPO, gestione sicura del procurement, gestione del personale, sicurezza delle infrastrutture fisiche, sicurezza delle operations, protezione, controllo degli accessi (sia a livello fisico che a livello software), protezione dai malware, logging e monitoring sicuro, network security, business continuity, compliance a leggi e regolamenti

# Budget, scala temporale

Le sezioni coinvolte nel progetto sono Perugia e CNAF.

PLANET è supposto durare due annualità ( 2021/22 ) e, come come da progetto iniziale, le richieste finanziarie sono **estremamente limitate:**

- Il primo anno solo missioni ( 3kEuro per sezione, assegnati ⅔ e restituiti tutti )
- Per il secondo anno chiediamo
  - **4kEuro missioni** ( reiteriamo la richiesta dell'anno scorso )
  - **2kEuro per contributo all'uso delle risorse EPIC** stimato assumendo cifre consip 2021 per cpu e 140E a TBN per il disco
    - Motivazione: uso "minimale" delle risorse della piattaforma EPIC per un anno per il trattamento dei dati ISS e, quando disponibili, i dati clinici.

NON è prevista NESSUNA richiesta per i servizi a PG



Daniele.Spiga@pg.infn.it

# Personale coinvolto e Anagrafica a CNAF **TEMPORANEA**

## Personale coinvolto:

- **PG**: **D.Spiga**, D. Ciangottini, P. Lubrano, M. Tracolli, L.Storchi, S.Cutini
  - Medici Associati: G. Ambrosio, F.Stracci, G.Reboldi
- **CNAF**: D.Salomoni, E.Ronchieri, B.Martelli, C.Duma, A.Costantini

| Nome | Percentuale |
|---|---|
| Costantini Alessandro | 10 |
| Duma Doina Cristina | 10 |
| Martelli Barbara | 10 |
| Ronchieri Elisabetta | 10 |
| Salomoni Davide (Resp. Locale) | 5 |

# Impatto e sinergie sviluppati nei primi mesi

- Insieme alla comunità dei Medici di PG abbiamo proposto un **PRIN** sul tema PLANET ( opportunamente esteso )
  - **UniPG Dipartimento di Medicina ( Capofila )** + INFN (PG e CNAF ) + UniMIB
    - Richesta 1MEuro

- Tirocinanti (**in Area Bologna/CNAF** ): già una Tirocinante ha lavorato con noi e potrebbero arrivarne altri

- La parte di “compute” della piattaforma di PLANET è stata **inclusa nel Progetto INFN-Cloud** come motore per la soluzione di Jupyter as a Service
  - E prima di questo la parte JupyterHub è stato adottato da INFN-Cloud come **soluzione base per il progetto ML-INFN ( CSN5 ) ed è quello che è stato utilizzato per** l’hackaton

- PLANET contribuisce ad EPIC attraverso la partecipazione allo sviluppo della versione sicura della piattaforma che verrà a breve dispiegata in EPIC
  - (in attesa di un passaggio di approvazione ufficiale che non dovrebbe presentare problemi)

# Qualche criticità c'è … ovviamente

- **Il recupero di dati:** Ottenere dati utili per uno studio "approfondito" è molto complicato e forse raggiungere la granularità geografica comunale che ci eravamo proposti potrebbe NON essere fattibile
  - Limite principale è il recupero di dati epidemiologici, per "motivi di privacy" (vedi i dati ISS)

- **La qualità dei dati**: la validazione di alcuni dati ( non ultimo quelli di inquinamento atmosferico ) non è affatto banale e probabilmente richiede esperti del settore
  - Qui abbiamo stabilito stretti contatti con ARPA Umbria da cui contiamo di ricevere supporto nell'interpretazione dei dati

- Infine: la Documentazione, come sempre siamo ancora un po' carenti..
  - Ma ci stiamo investendo … Abbiamo iniziato a creare una documentazione per l'accesso ai dati e per l'amministrazione del sistema che stiamo sviluppando e proponendo Confluence

# Alcune note finali

**L'attività è molto interessante per tanti motivi diversi:**

- "Scientifico": estremamente multidisciplinare e tanto scambio di conoscenze tra comunità diverse, quali Epidemiologi/Clinici, Fisici/Analisti, Computing e Trattamento dati (più o meno sensibili)
  - Infatti abbiamo da subito parlato di "Data Lake" model / Piattaforma di calcolo; EPIC etc..
  - NON Banale far parlare tutti la stessa lingua!

- "Tecnologico": Per noi la finalità non è SOLAMENTE l'analisi sull'associazione statistica
  - Vogliamo anche impostare un modello riusabile, estendibile e adottabile a problematiche diverse
    - dalla gestione di set di dati eterogenei, alle tecniche di analisi passando per le norme del GDPR (**EPIC**)

Tutto questo rende il progetto un po più impegnativo del previsto, ma sicuramente ci permette di imparare molte cose e sviluppare molte soluzioni riusabili all'interno ente!

- L'interazione tra le due anime (clinico/epidemiologica e INFN) cresce sempre più e questo ha un impatto importante sulla produttività e sui risultati che si possono ottenere

# Backup



Daniele.Spiga@pg.infn.it

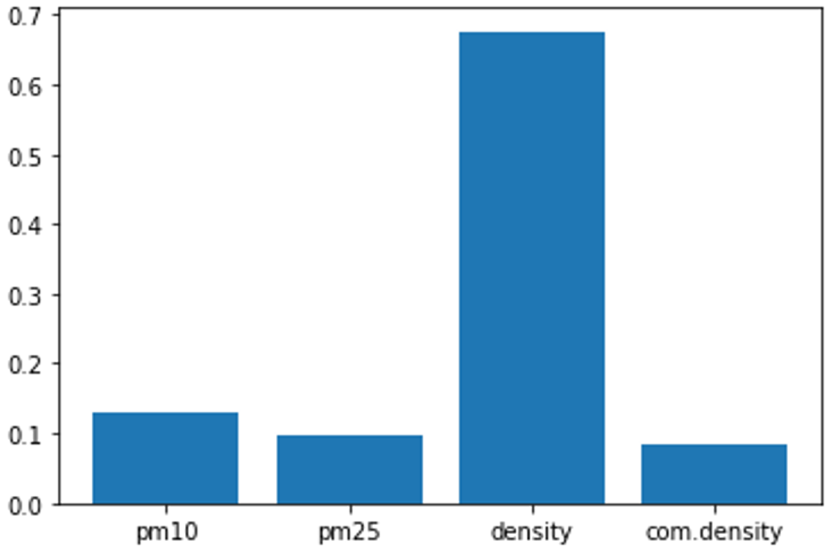
# Analisi dei dati VERY PRELIMINARY

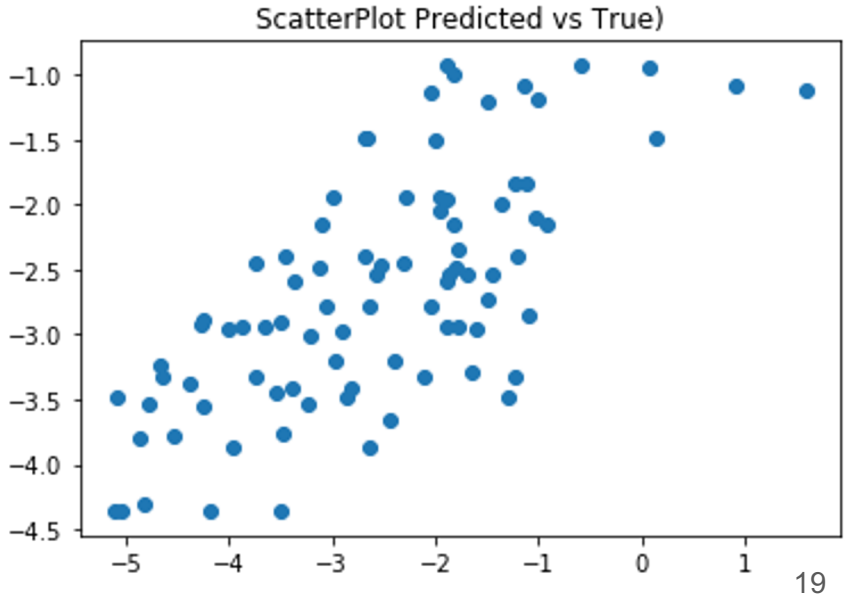
**Abbiamo appena avviato una analisi temporale** usando dati nazionali

- KNeighborsRegressor e **Permutation feature importance** per valutare l'importanza delle features

Dataset:
Copernicus 2017 - 2020

Dati Covid19: da
Setti L, ey.al. *BMJ Open* 2020;**10:**e039338






Daniele.Spiga@pg.infn.it