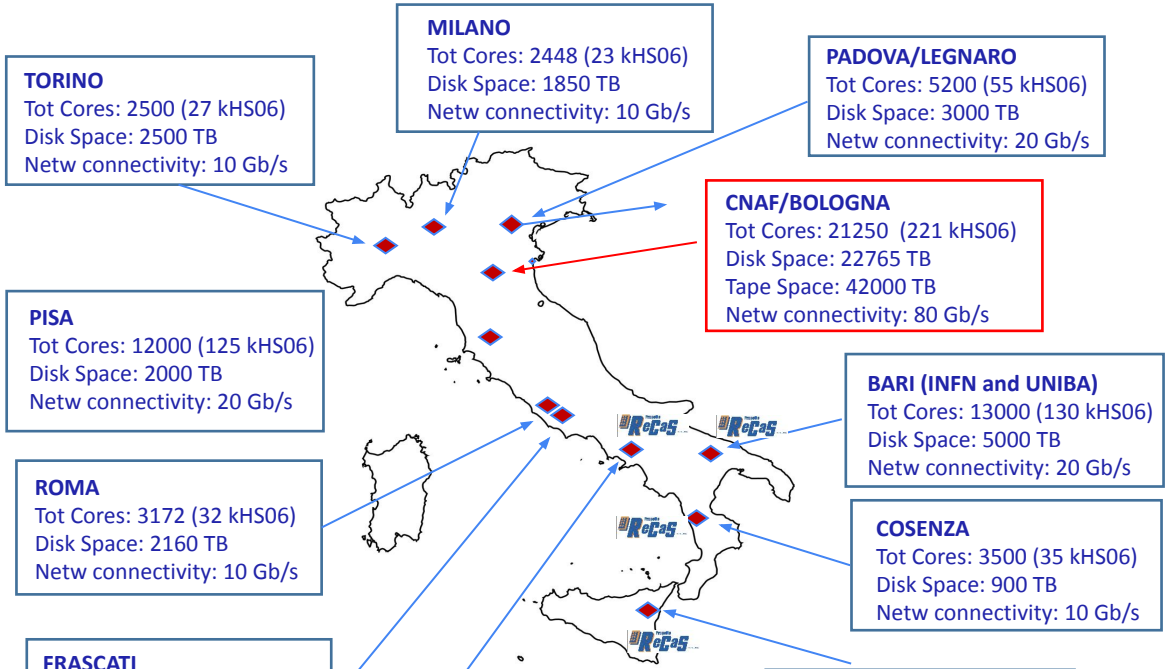
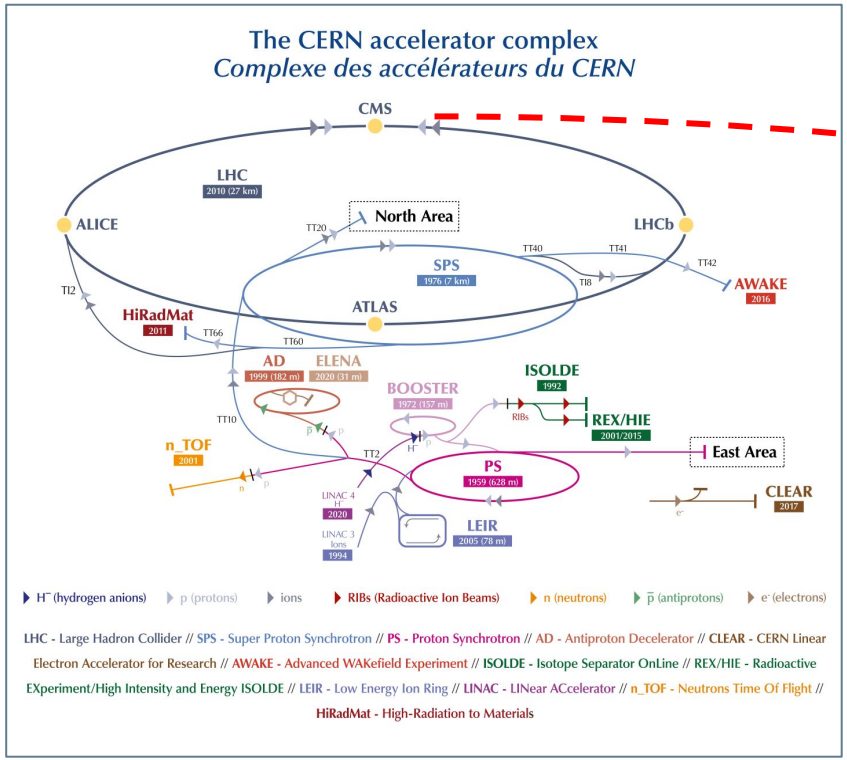
Computing / Data Science: Attività di ricerca tecnologica nel campo del calcolo scientifico

Daniele Spiga

INFN Scientific computing facilities



Computing, software and data science nella fisica: Partiamo dall'esempio di LHC



- Le particelle in collisione creano nuove particelle, i cui prodotti di decadimento fluiscono attraverso strati di rivelatori

Da dove vengono i dati… e dove vanno ?

- I rivelatori producono segnali elettrici in risposta ai flussi di particelle
- Le informazioni provenienti da ogni parte del rilevatore vengono combinate per creare un riepilogo digitale dell '"evento di collisione".

Ma poi ci sono anche i dati simulati dai ricercatori: “la verità MonteCarlo”

I segnali elettrici rilasciati dalle particelle negli apparati vengono digitalizzati e scritti su Hard Disk

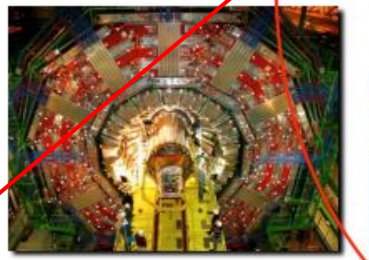
daniele.spiga@pg.infn.it

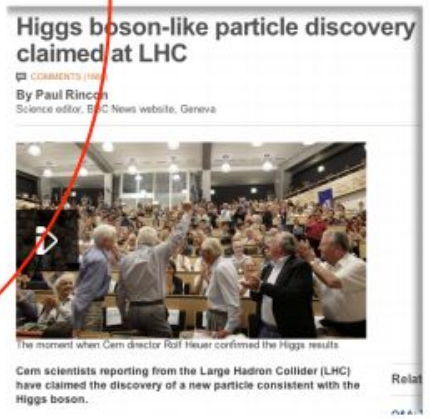
Quindi come si arriva ad una scoperta

Acceleratore ➡ Experimento ➡ Analisi dei **Dati** (Computing) Grid / Cloud Big Data ➡ Scoperta

Produce **Dati**

"Fotografa" **Dati**





Oggi ci concentriamo su questo

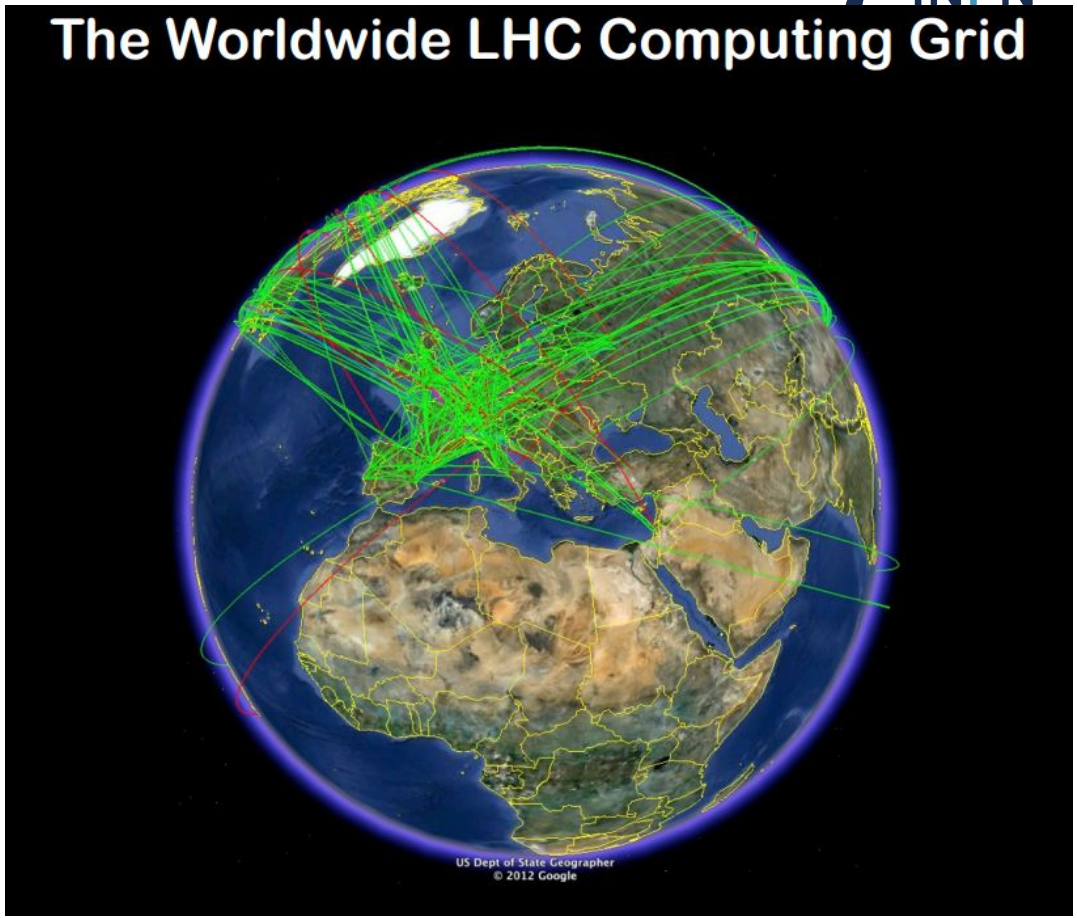
Cosa è stato fatto negli ultimi 20 anni.

La Grid:

> **“Coordinated resource sharing and problem solving in dynamic, multi institutional virtual organizations”**
>
> Ian Foster e Karl Kesselman

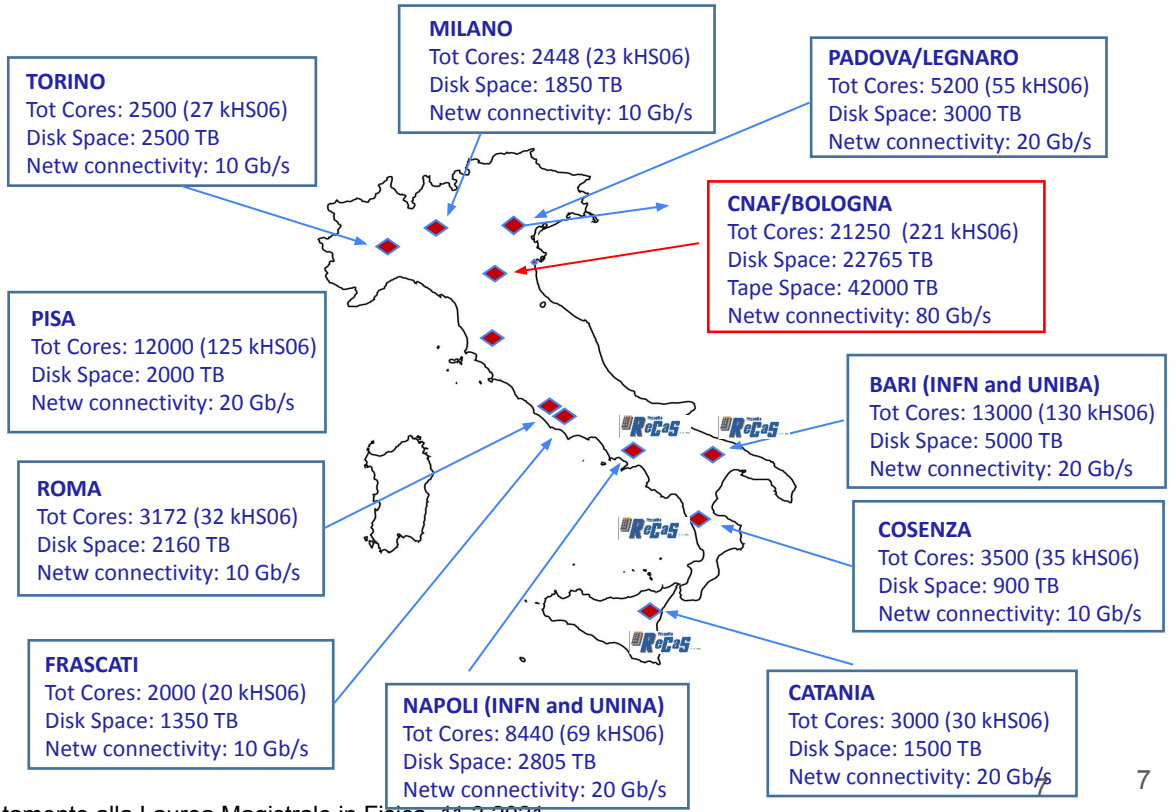
Dal punto di vista dell’utente:

- Io voglio usare le risorse di calcolo ogni volta che ne ho bisogno
- Non mi interessa chi è Il proprietario o dove sono
- I miei programmi devono girare sulle risorse disponibili.



INFN Scientific computing facilities

- **INFN ~ 10% of WLCG for 2017**
  - 15-20% ALICE
  - ~13% CMS
  - ~7-10% ATLAS
  - ~10-12% LHCb



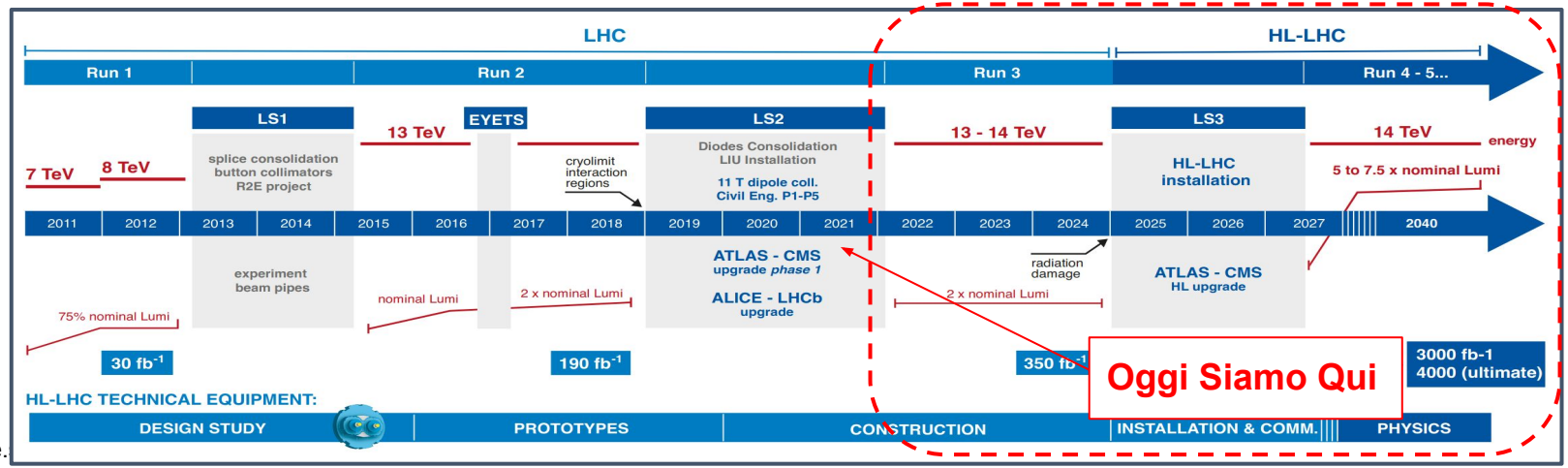
WLCG A Success Story

CERN July 4th 2012

Quindi, è tutto fatto?

NO!! Per quanto riguarda l'evoluzione degli esperimenti al CERN si va verso HL-LHC.

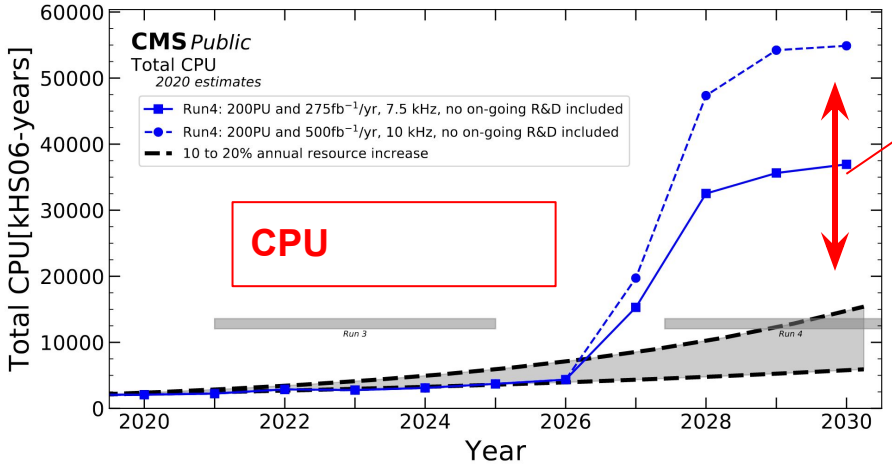
- Questo significa un acceleratore più potente, e dei rivelatori capaci di immagazzinare più quantità di dati (fare più fotografie)
- Questo pone nuove sfide di fisica ma porta sfide notevoli anche dal punto di vista della **mole di dati da immagazzinare e da processare**

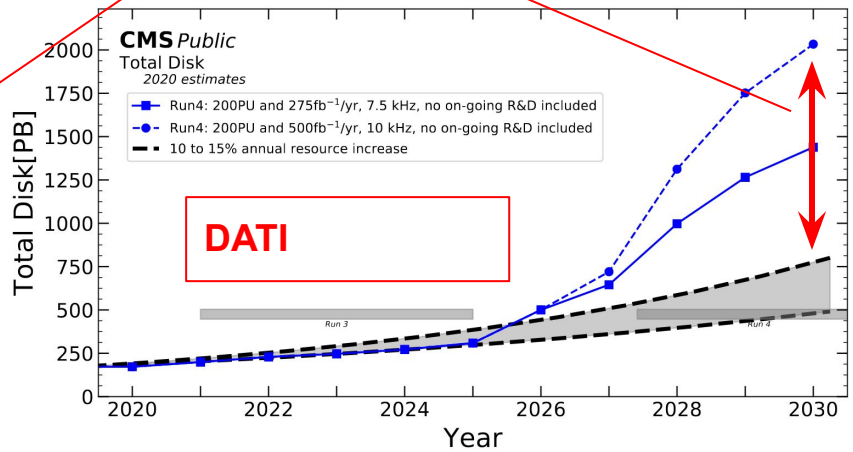


Quantifichiamo un pochino…

Questo è il problema computazionale da risolvere

Punto di vista CMS:



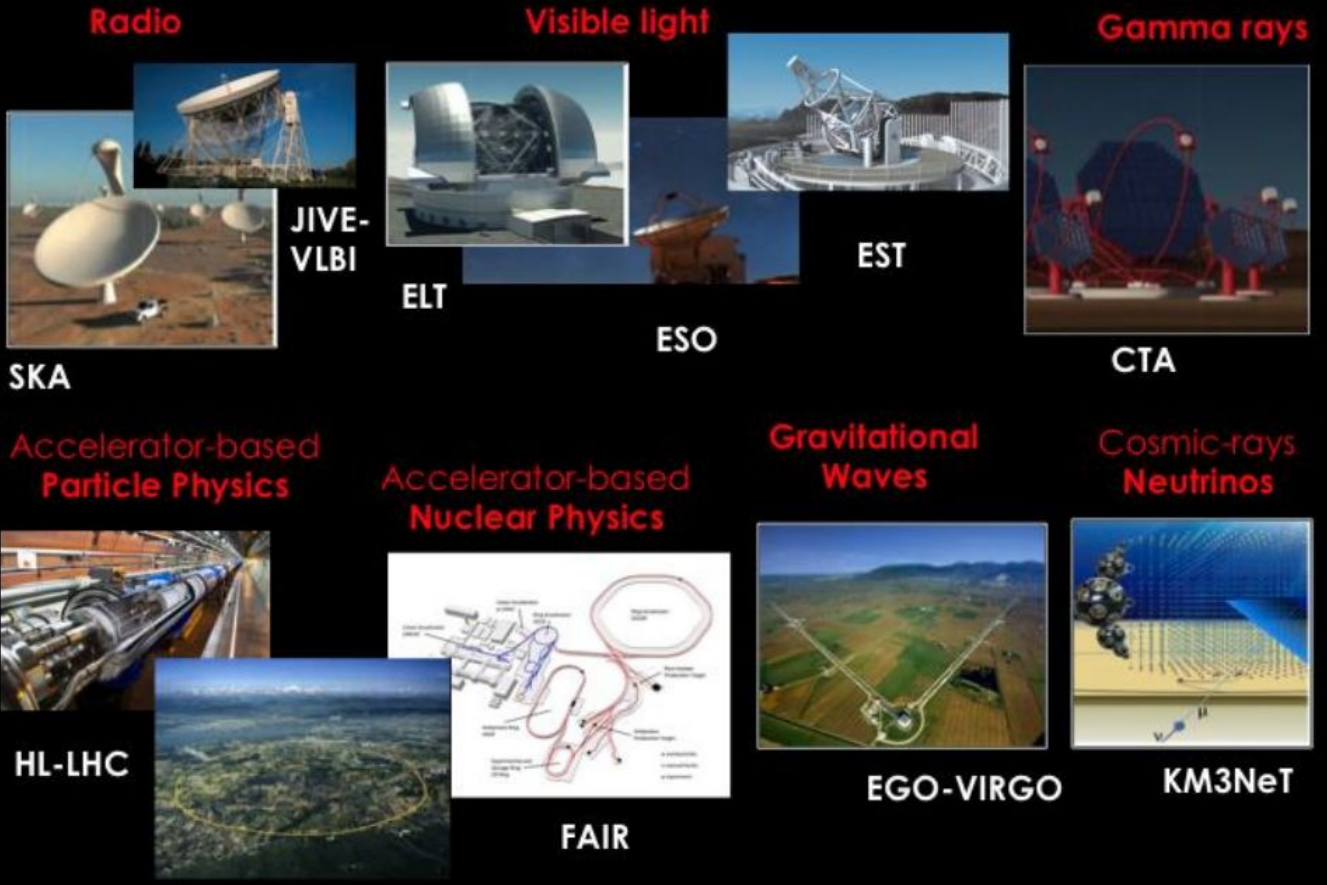


CPU: **3x rispetto alla proiezione**.

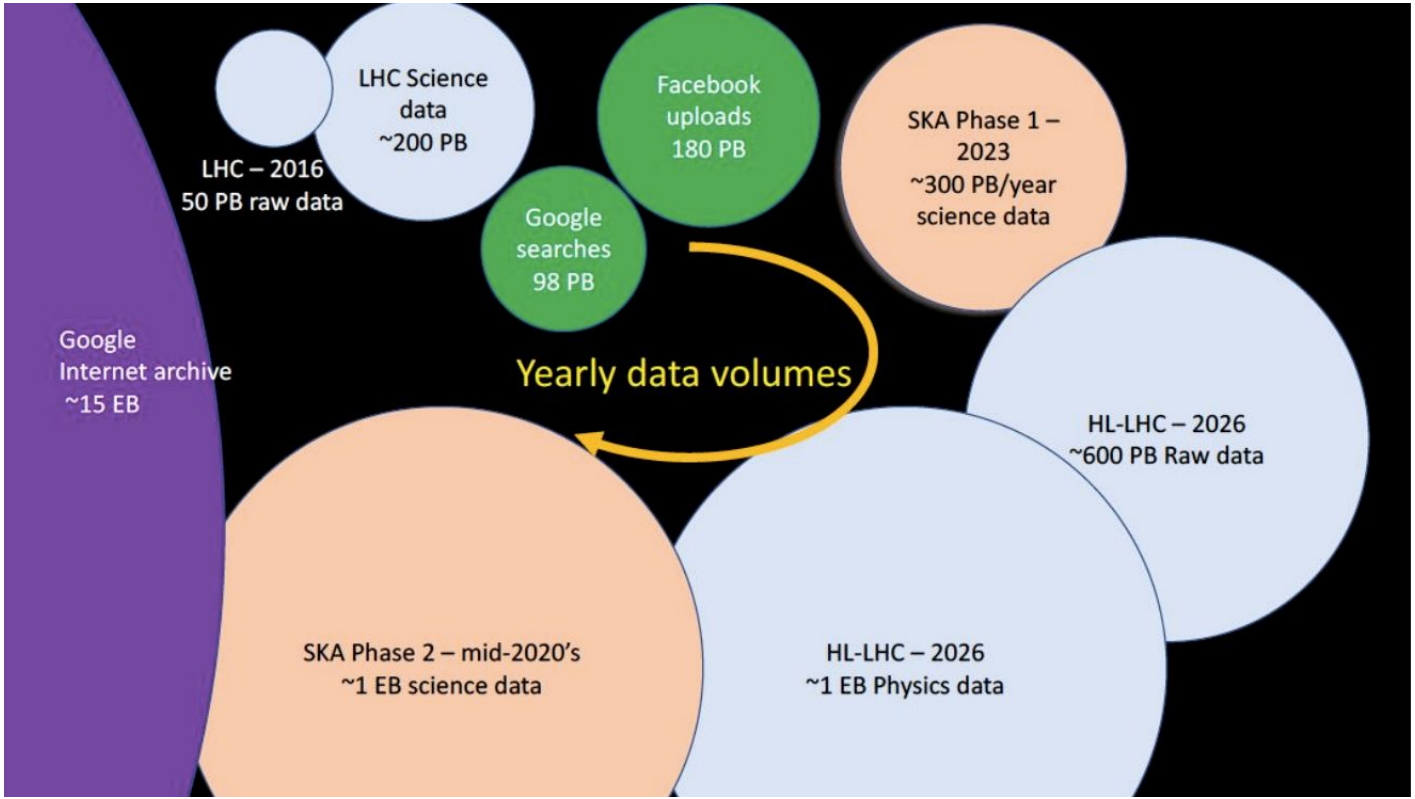
- NOTA: questo conto non incorpora l'effetto degli acceleratori

Esigenze disco: **2.5x rispetto alla disponibilità prevista**

Ma NON solo CERN ed LHC, tanti altri esperimenti….



Un po di numeri a confronto



e non solo fisica… i dati eterogenei (Data Science)

ISTAT
ISTAT Censimento
ID (Dr. N. Caranci)
Dati covid Regione Umbria
ARPA Umbria

Partendo dai **dati RAW** raccolti e archiviati nel formato originale possiamo generare datasets integrati funzionali all'analisi

- primo dataset integrato per avviare la fase di "validazione" ed esplorazione

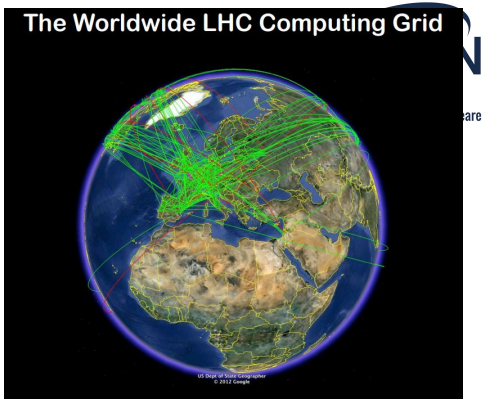
	City	Population	Density	Surface	Lattante	Primainfanzia	Secondainfanzia	Terzainfanzia	Adolescenza	Primaadulta	Secondaadulta	Terza
0	Acquasparta	4611	57.0	81.61	48.0	22.0	150.0	222.0	484.0		1368.0	915.0
1	Allerona	1722	21.0	82.61	20.0	13.0	53.0	63.0			503.0	370.0

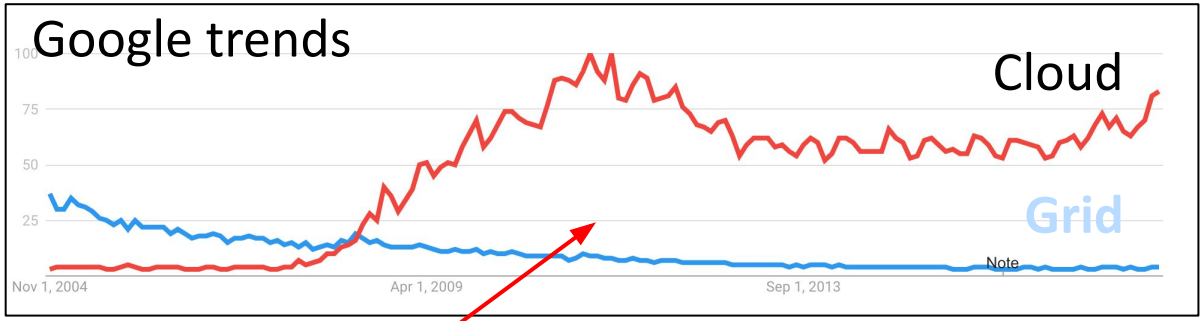
Quarta	Quinta	TotaleEta	MediaEta	Depriv_idx	Depriv_cat			Deceased	MaxIntensiveCare	AvgIntensiveCare
579.0	74.0	4570.0	47.441357	1.0		3	0.380952	2	1	0.027211
207.0	29.0	1724.0	48.408353	-1.3		41	2	0.262238	0	0.000000

COVID-19 e Inquinamento Atmosferico

mean_pm10_ug/m3_mean_2019	mean_pm10_ug/m3_std_2019	mean_pm10_ug/m3_median_2019	mean_pm10_ug/m3_mean_2020	mean_pm10_ug/m3_std_2020
16.183606	7.358666	14.376109	15.833302	6.970052
15.828832	7.167570	14.451391	14.905208	7.113094

Intanto: l'evoluzione tecnologica

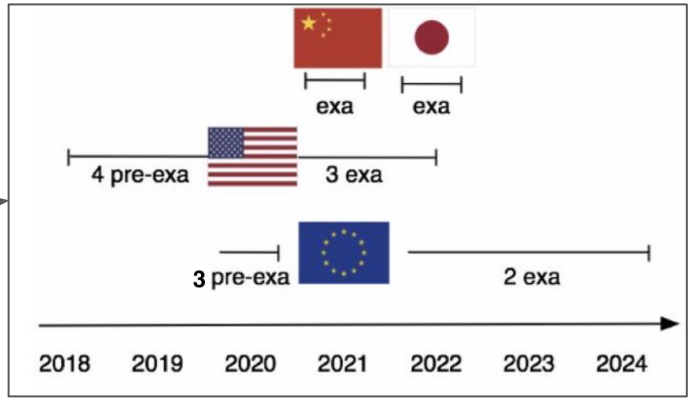




Cloud Computing

High Performance Computing

- **GPU, FPGA, PowerPC..**



Quindi, cosa si fa? Quali settori di ricerca?

Istituto Nazionale di Fisica Nucleare

Cosa facciamo a Perugia

Si fa ricerca tecnologica per testare e studiare

nuove infrastrutture di calcolo, come **Cloud e HPC** per costruire soluzioni di :

- **Analysis Facility**, **Datalake**

Si sviluppano di nuovi algoritmi per trovare soluzioni intelligenti basati su tecniche di:

- **Machine Learning** e quindi --> **Intelligenza artificiale**

Si integrano tecnologie per la gestione, il trattamento e l'analisi di dati eterogenei applicando soluzioni di :

- **Data Science**

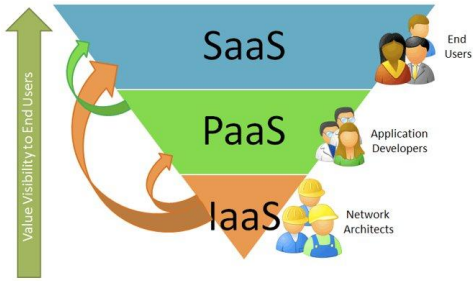
daniele.spiga@pg.infn.it

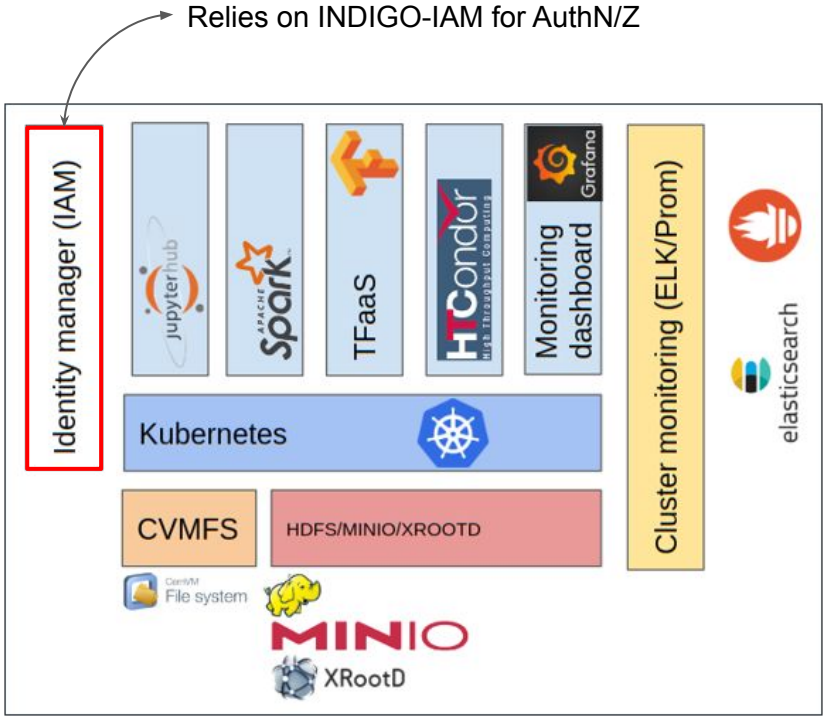
Nuove infrastrutture

Fornire una struttura a blocchi (i.e Mattoncini Lego) flessibile a livello PaaS, basata su standard aperti dell'industria per seguire il paradigma "service composition model"

- per supportare workflow scientifici su grandi quantità di dati (Big Data) sfruttando una Cloud ibrida

- ❏ User oriented
- ❏ Highly customizable
- ❏ Community agnostic





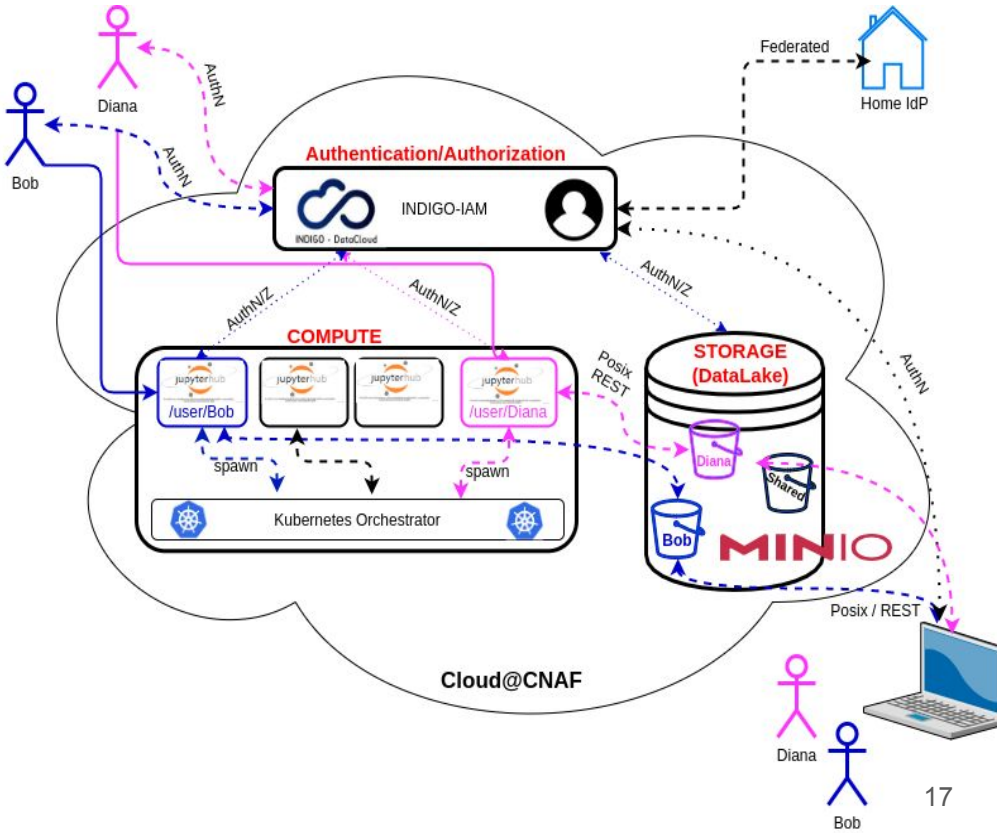
The DODAS lego blocks ecosystem

Fully compatible with Cloud orchestrator engines, such as the INDIGO-PaaS Orchestrator

Il Data Lake

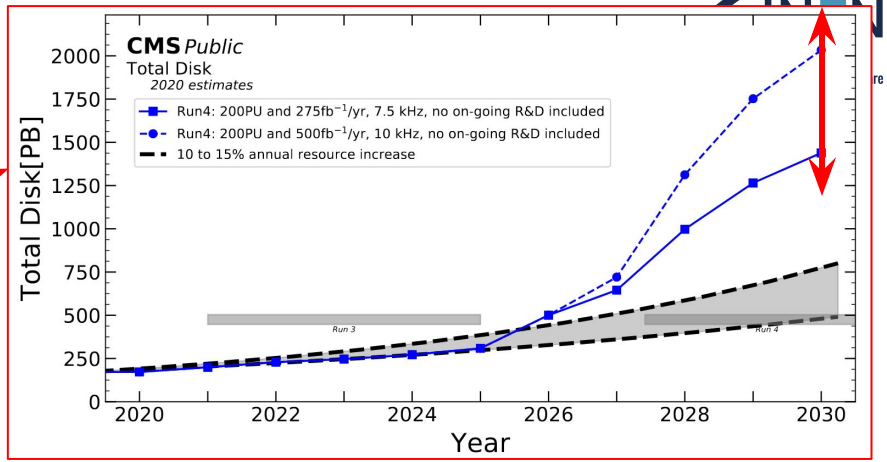
Integrare e rendere disponibile una piattaforma “open” e generica (riutilizzabile):

- Integrazione di dati da sorgenti informative multiple
- Processamento (analisi descrittiva, predittiva e real time)
- Federazione di risorse di calcolo attraverso la tecnologia abilitante di INFN-Cloud



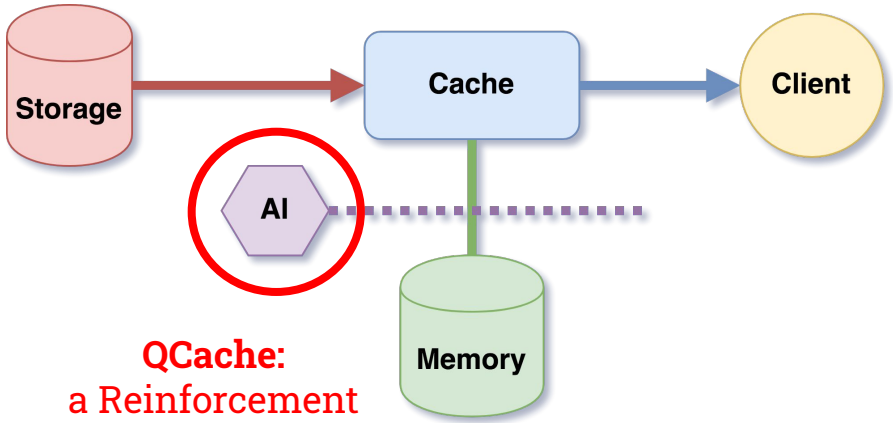
Intelligenza Artificiale

Per risolvere questo



Progetto per creare un sistema di cache intelligente tramite l'utilizzo di algoritmi di IA nel contesto dell'esperimento CMS:

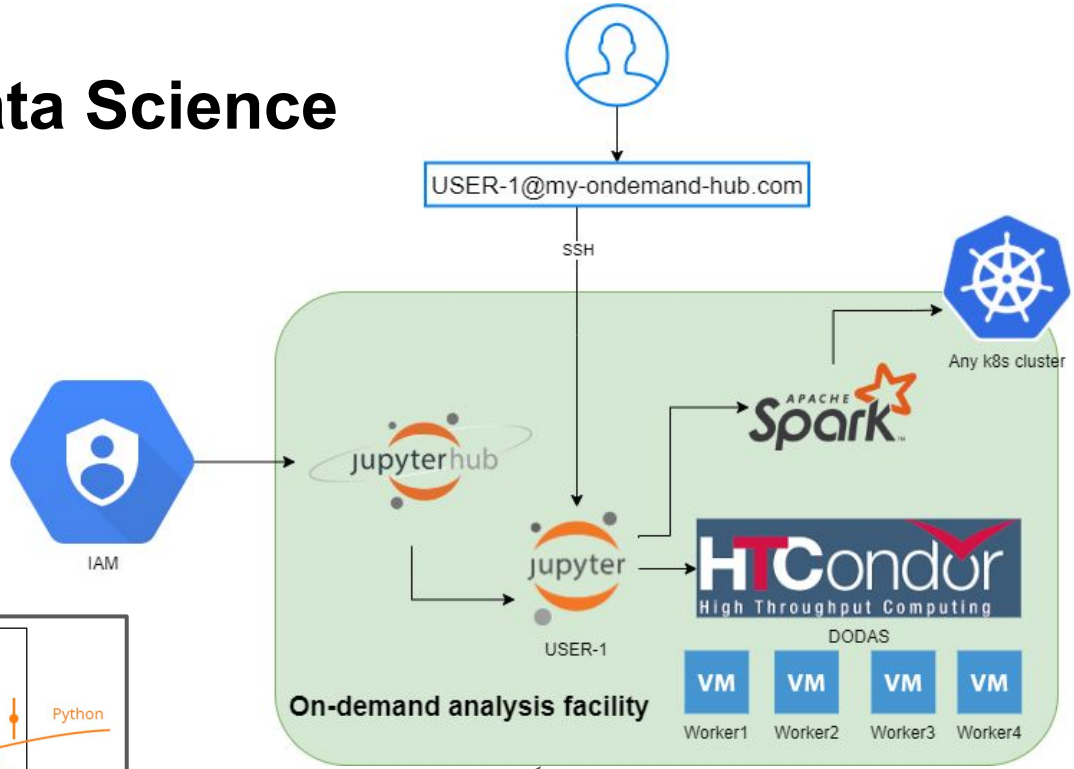
- L'IA manipola la memoria della cache, decidendo cosa scrivere o cancellare
- Il risultato che si vuole ottenere è quello di avere un algoritmo che usa meno risorse di storage in confronto agli algoritmi classici (LRU, LFU, etc..), mantenendo performance comparabili

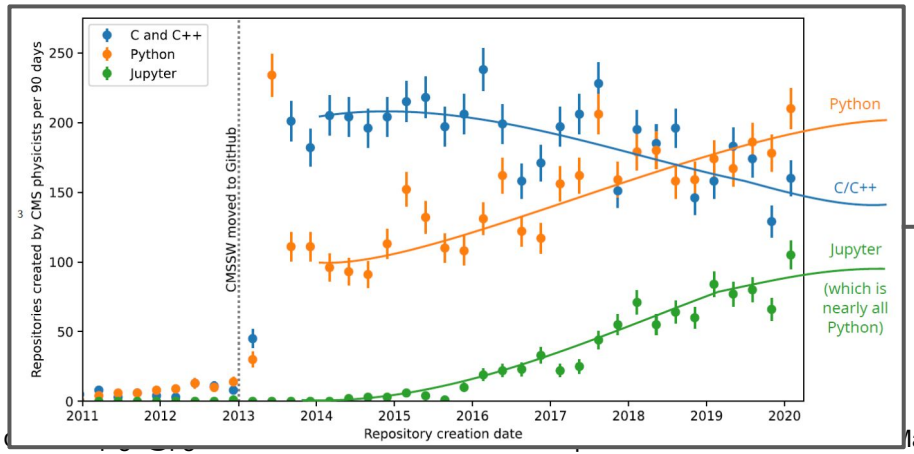


QCache: a Reinforcement Learning-based framework

Analysis Facility per Data Science

- Analisi interattiva tramite notebook jupyter
- Utilizzo di tecnologie standard per l'attuale approccio all'analisi dati in ambito "data science"





"software di CMS sempre più pythonico"

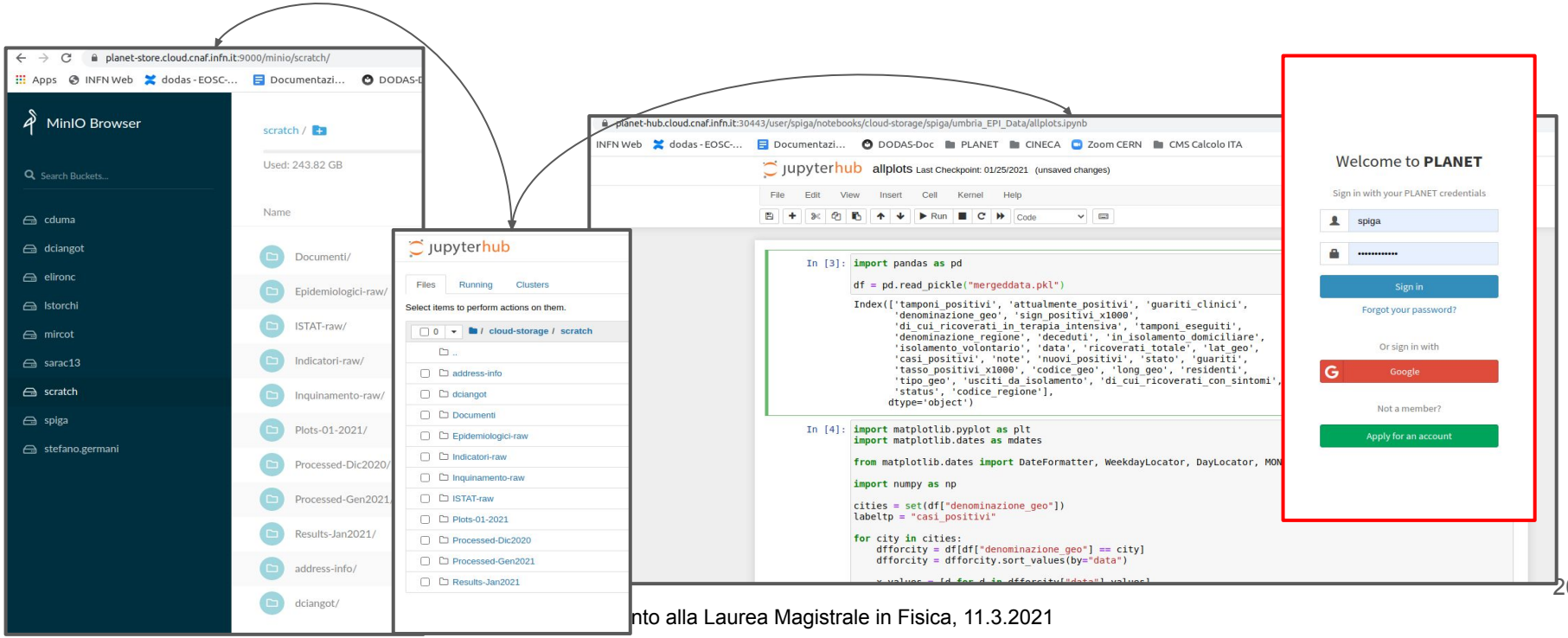
quindi perfetto per questo approccio..

Analisi dati eterogenei con tecniche di Data Science

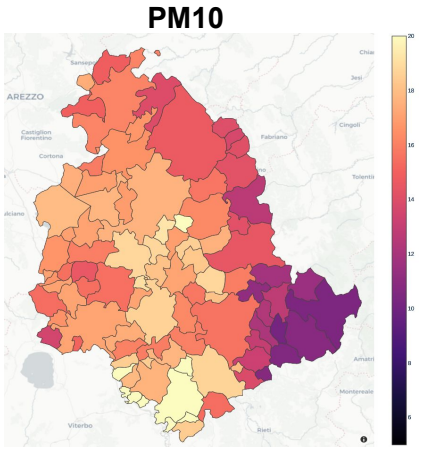
Istituto Nazionale di Fisica Nucleare

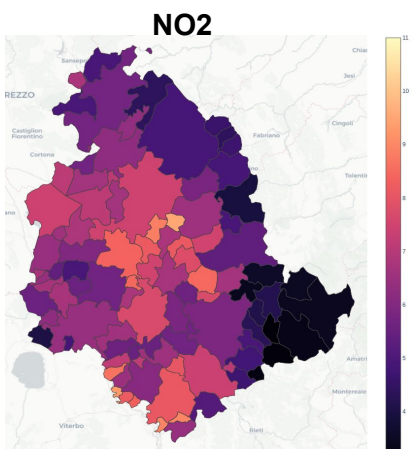
Integrare e rendere disponibile una piattaforma “open” e generica (riutilizzabile) per

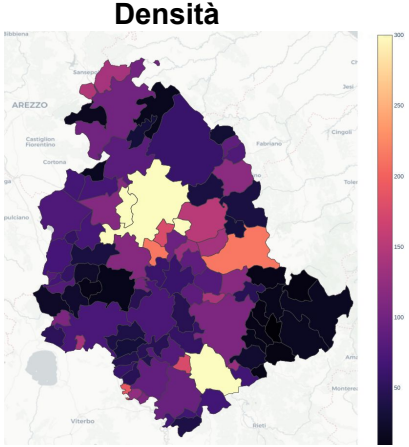
- Integrazione di dati da sorgenti informative multiple, Processamento, Federazione di risorse attraverso soluzioni di **INFN-Cloud**

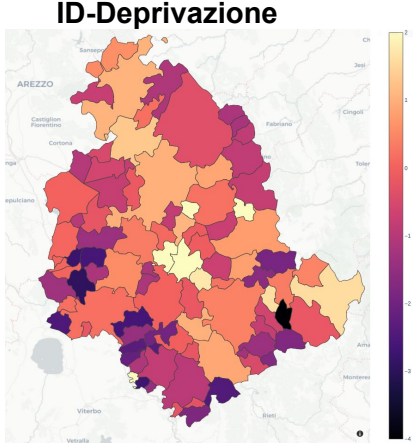


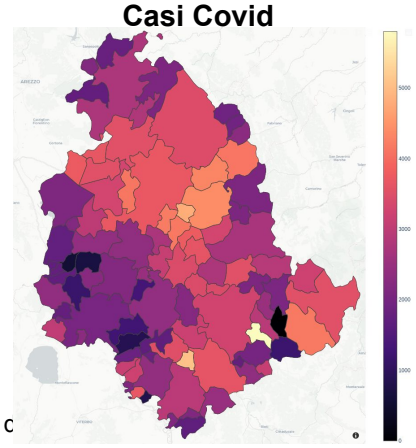
Esempio: Covid-19 e inquinamento in Umbria

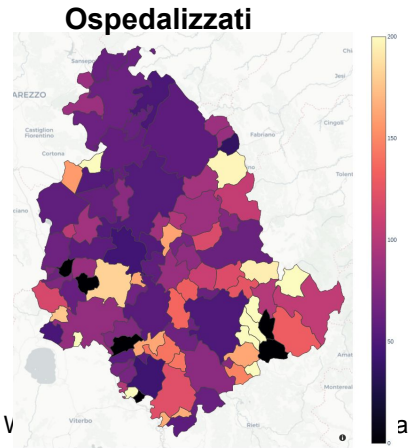


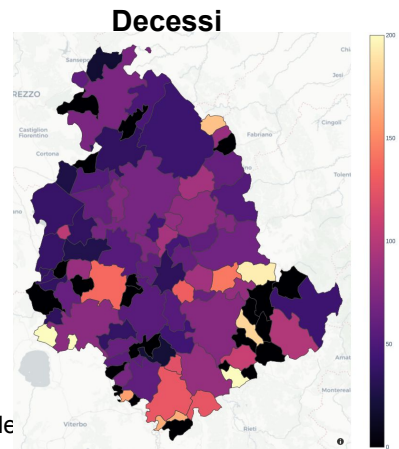


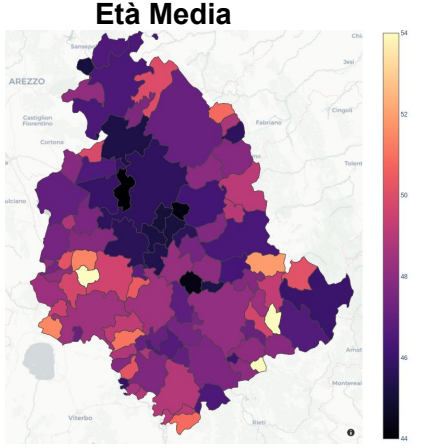












C V a Magistrale

Riassumendo quali attività di tesi:

Il Calcolo Scientifico offre molte opportunità di ricerca in settori all'avanguardia. In particolare a Perugia:

- **Cloud Computing**:
  - Sviluppo di servizi per l'analisi dei dati ad LHC
  - Integrazione di HPC nel modello di calcolo di CMS
  - Sviluppo di soluzioni di Data Lake per la gestione di dati eterogenei

- **Intelligenza Artificiale:**
  - Sviluppo di sistemi intelligenti per la gestione dei bigdata
  - Sviluppo di algoritmi per l'ottimizzazione della selezione dei dati nella fisica di CMS
    - signal vs background discrimination for Vector Boson Scattering same sign WW with hadronic tau decay

- **Data Science:**
  - Progetto PLANET per lo studio dell'associazione Covid-19 e Inquinamento Atmosferico

Contatti **daniele.spiga@pg.infn.it**