

LAURA REDAPI

Consortium
GARR

THE ITALIAN
EDUCATION
& RESEARCH
NETWORK



HEPMED: data mining in **H**igh **E**nergy **P**hysics and **MED**icine

GIORNATA DI INCONTRO
BORSE DI STUDIO GARR
"ORIO CARLINI"
ROMA

Roma, 27 Giugno 2019

Borsisti Day 2019



Sviluppo di un modo efficiente di navigare grandi quantità di dati sfruttando tecniche di data warehousing, con applicazione diretta su database di **fisica delle alte energie** e di ambito **medico**

Abbiamo a disposizione dati di misure fatte dagli anni '70 ad oggi potrebbe esserci sfuggito qualcosa?



Correlando dati di esami medici diversi, potrei forse fare diagnosi più sicure?





IDEA

Creare uno strumento per rispondere alla domanda:



I dati sono consistenti con la teoria XXX?



Di cosa abbiamo bisogno?:

1) Un modo efficace per navigare i dati

- Organizzare i **dati già disponibili** ed estrarre **automaticamente** quante **più informazioni possibili** tramite un **interfaccia** di supporto

Quali sono le misure rilevanti per la teoria da testare?

Vorrei confrontare dati di analisi diverse su un unico grafico...

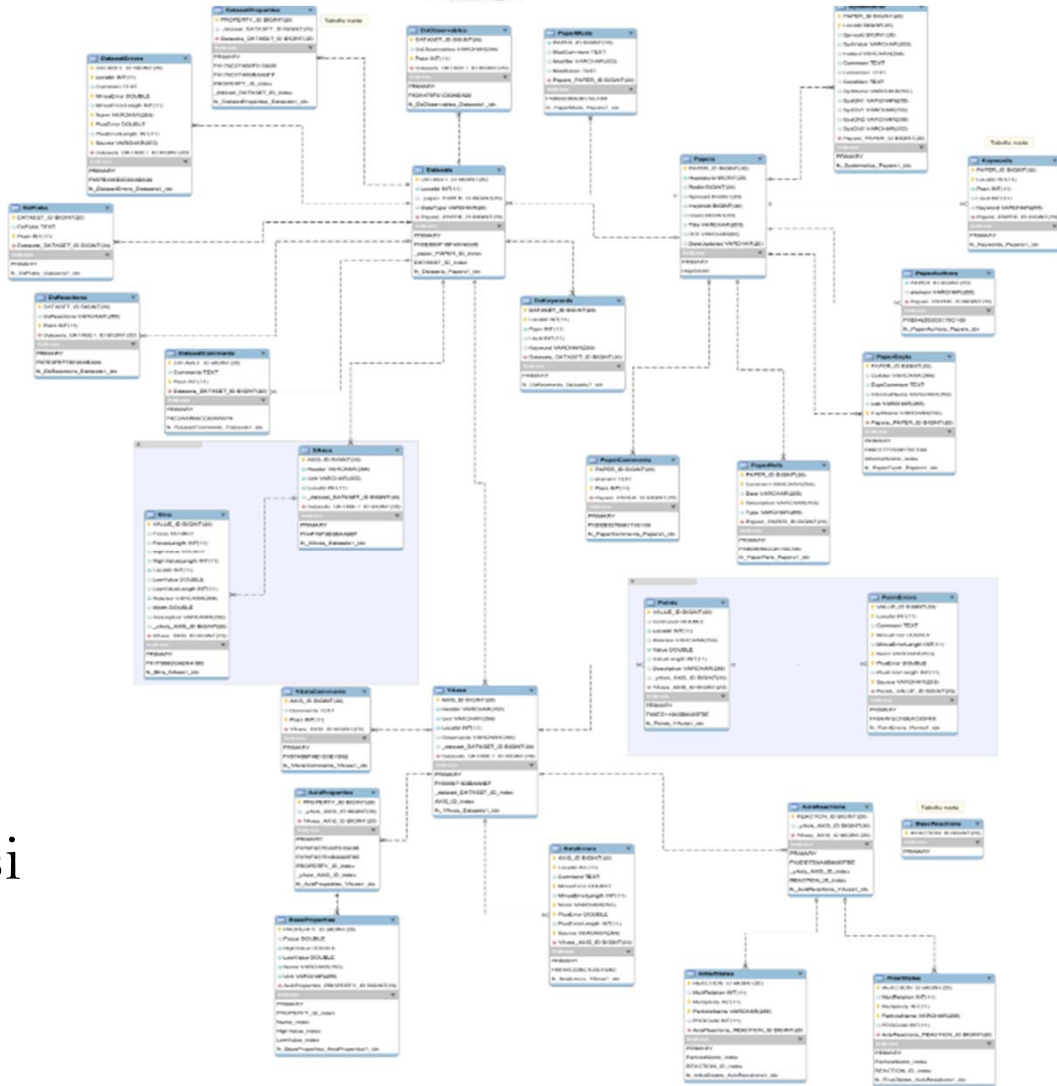
2) Un tool per ottenere previsioni teoriche delle quantità misurate

3) Uno **strumento statistico** per quantificare l'accordo tra dati e teoria



Start-point 1: database hepdata

- Database relazionale con struttura complessa
 - Elemento base della ricerca è l'articolo scientifico
- Ricerco tutti quei papers che...*
- Possibilità di navigare i dati all'interno di una singola analisi selezionata

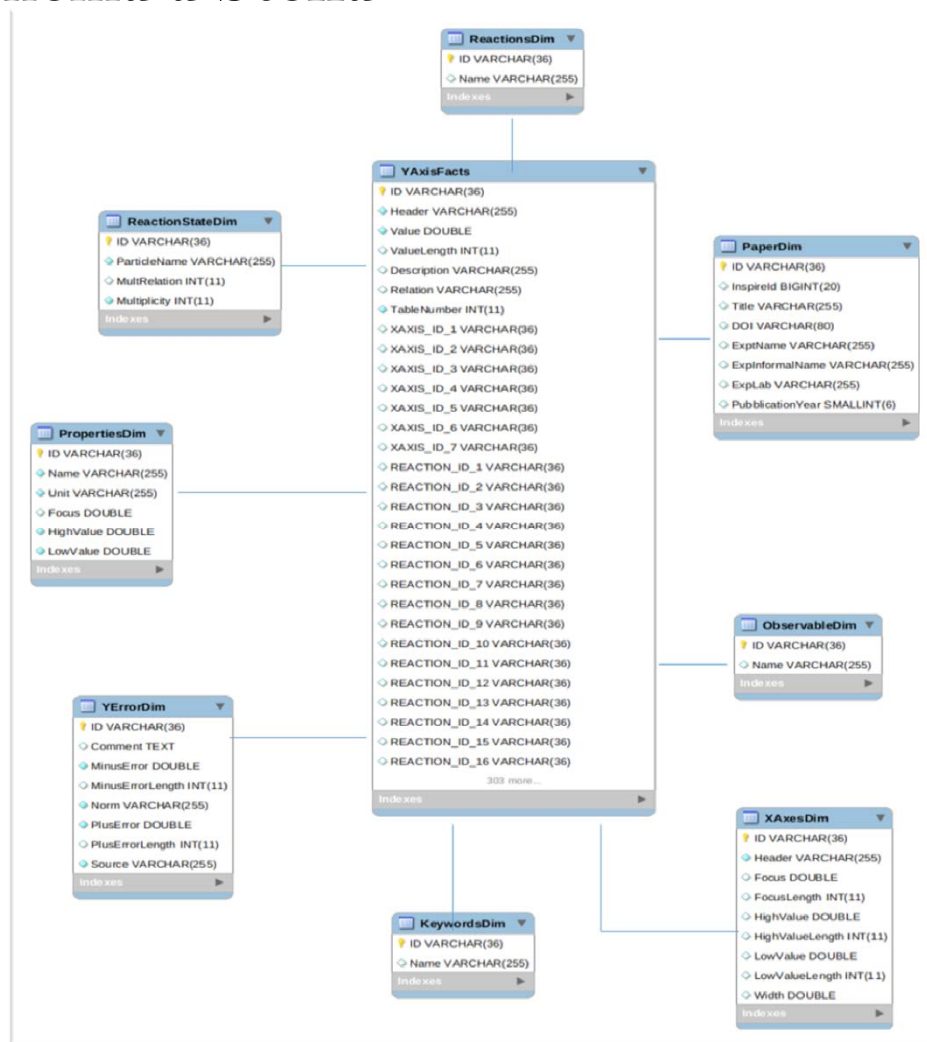




Il Nuovo Database MineHep



Schema a Stella

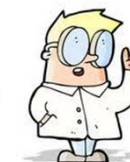


YAxisFact → dato numerico della singola misura

Le **dimensioni** identificate sono :

- Properties
- YError
- XAxes
- Observable
- Paper
- Keywords
- Reaction

Ogni dimensione è una tabella con tante colonne quante sono le caratteristiche necessarie ad identificarla.



Test di controllo

-- TEST: prendo tutte le misurazioni della tabella 1 che sono comprese tra [0.8, 2.0] η nella colonna HERAPDF

```
select distinct Value
from YAxisFacts yaf
inner join PaperDim ppd on ppd.ID = yaf.PAPER_ID
inner join XAxisDim xad on xad.ID = yaf.XAXIS_ID_1
where InspireId = 1118047 and TableNumber = 1
      and xad.Header = '$|\\eta|'
      and yaf.Header = "HERAPDF"
      and xad.LowValue >= 0.8
      and xad.HighValue <= 2.0;
# Value
# 196
# 181
# 153
# 140
# 132
```

-- RESULT: OK

-- TEST: prendo tutte le misurazioni della tabella 1 della colonna A che hanno un errore SYS +/- 6

```
select yaf.Value
from YAxisFacts yaf
inner join PaperDim ppd on ppd.ID = yaf.PAPER_ID
inner join YErrorDim yed on yed.ID = yaf.POINT_ERROR_ID_1
where InspireId = 1118047
      and TableNumber = 1
      and Source = 'SYS'
      and Header = '$|\\mathcal{A}|'
      and (MinusError=-6 and PlusError=6);
# Value
# 156
# 136
```

-- RESULT: OK

PT(E)	> 35 GEV				
RE	P P -> W+ -< E+ - NUE > X				
SQRT(S)	7000.0 GeV				
$ \eta $	$\mathcal{A}$	CT10	HERAPDF	MSTW	NNPDF
0.0 - 0.2	102.0 ± 3.0 stat ± 5.0 sys	109.0 ± 5.0	106.0 $^{+4.0}_{-4.0}$	87.0 $^{+3.0}_{-3.0}$	107.0 ± 5.0
0.2 - 0.4	111.0 ± 3.0 stat ± 5.0 sys	114.0 ± 5.0	110.0 $^{+4.0}_{-4.0}$	89.0 $^{+3.0}_{-3.0}$	110.0 ± 5.0
0.4 - 0.6	116.0 ± 3.0 stat ± 5.0 sys	119.0 ± 5.0	115.0 $^{+4.0}_{-4.0}$	98.0 $^{+3.0}_{-3.0}$	116.0 ± 5.0
0.6 - 0.8	123.0 ± 3.0 stat ± 5.0 sys	126.0 ± 5.0	122.0 $^{+4.0}_{-4.0}$	103.0 $^{+3.0}_{-3.0}$	123.0 ± 5.0
0.8 - 1.0	133.0 ± 3.0 stat ± 5.0 sys	138.0 $^{+5.0}_{-6.0}$	132.0 $^{+4.0}_{-4.0}$	115.0 $^{+4.0}_{-3.0}$	134.0 ± 5.0
1.0 - 1.2	136.0 ± 3.0 stat ± 6.0 sys	146.0 ± 6.0	140.0 $^{+5.0}_{-6.0}$	128.0 $^{+4.0}_{-3.0}$	145.0 ± 5.0
1.2 - 1.4	156.0 ± 3.0 stat ± 6.0 sys	164.0 $^{+6.0}_{-7.0}$	153.0 $^{+5.0}_{-7.0}$	144.0 ± 5.0	158.0 ± 5.0
1.6 - 1.8	166.0 ± 3.0 stat ± 10.0 sys	195.0 $^{+8.0}_{-9.0}$	181.0 ± 5.0	179.0 ± 5.0	190.0 ± 4.0
1.8 - 2.0	197.0 ± 3.0 stat ± 9.0 sys	207.0 $^{+8.0}_{-10.0}$	196.0 $^{+4.0}_{-3.0}$	200.0 $^{+6.0}_{-3.0}$	206.0 ± 4.0
2.0 - 2.2	224.0 ± 3.0 stat ± 11.0 sys	224.0 $^{+8.0}_{-11.0}$	211.0 $^{+5.0}_{-3.0}$	213.0 $^{+6.0}_{-3.0}$	219.0 ± 4.0
2.2 - 2.4	210.0 ± 4.0 stat ± 13.0 sys	241.0 $^{+8.0}_{-12.0}$	225.0 $^{+8.0}_{-4.0}$	231.0 $^{+6.0}_{-3.0}$	231.0 ± 5.0



State of the art: MINEHEP + METABASE



Metabase, strumento opensource, per interrogare ed estrarre conoscenza dai dati in modo semplice tramite un'interfaccia SQL

Metabase interface showing a query and its results.

Database: minehep-bridgetables

Query:

```
SELECT distinct PaperDim.InspireId, yaf.TableNumber, PropertiesDim.Name, PropertiesDim.LowValue, PropertiesDim.HighValue
FROM YAxisFacts yaf
JOIN PropertiesBridge ON yaf.BRIDGE_PROPERTY_ID = PropertiesBridge.GROUP_ID
JOIN PropertiesDim ON PropertiesBridge.PROPERTY_ID = PropertiesDim.ID
INNER JOIN PaperDim ON PaperDim.ID = yaf.PAPER_ID
JOIN XAxesBridge ON XAxesBridge.GROUP_ID = yaf.BRIDGE_XAXIS_ID
JOIN XAxesDim ON XAxesBridge.XAXIS_ID = XAxesDim.ID
JOIN ReactionsBridge ON ReactionsBridge.GROUP_ID = yaf.BRIDGE_REACTION_ID
JOIN ReactionsDim ON ReactionsDim.ID = ReactionsBridge.REACTION_ID
JOIN KeywordsBridge ON KeywordsBridge.GROUP_ID = yaf.BRIDGE_KEYWORD_ID
JOIN KeywordsDim ON KeywordsBridge.KEYWORD_ID = KeywordsDim.ID
JOIN ObservablesBridge ON ObservablesBridge.GROUP_ID = yaf.BRIDGE_OBSERVABLE_ID
JOIN ObservableDim ON ObservableDim.ID = ObservablesBridge.OBSERVABLE_ID
```

Visualizzazione: Tabella

Inspire ID	Table Number	Name	Low Value	High Value
883,318	2	SQRT(S)	7,000	7,000
1,089,993	2	SQRT(S)	7,000	7,000
1,240,670	5	SQRT(S)	7,000	7,000
1,238,809	5	SQRT(S)	7,000	7,000





Risultati ottenuti :



- Elemento principale della ricerca è il **dato** della singola analisi
- Possibilità di fare **query articolate**

Cercare tutti i dati che rispondono a certe keywords e a tagli su diverse proprietà scelte dall'utente

- Possibilità di **navigare su più analisi** contemporaneamente:

Osservo tutte le tabelle di interesse contemporaneamente e decido quali dati, anche di esperimenti diversi, riportare in uno o più grafici



Step futuri:

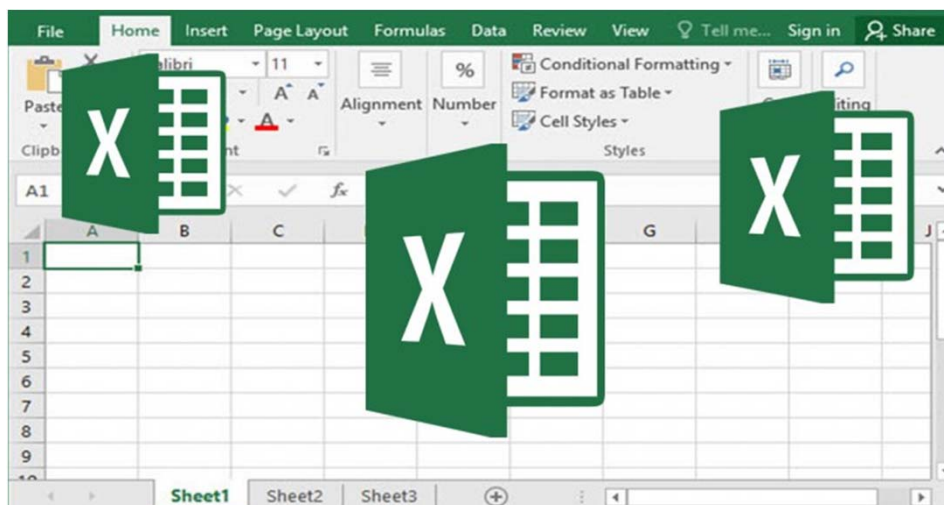
- Vocabolario (eterogeneità del dato)
- Interfaccia ad hoc (linguaggio php e python)



Start-point 2: database fisica medica



- Verifiche pretrattamento della radioterapia dell'Azienda Ospedaliera Universitaria di Careggi
- Risultati analisi gamma
- Fogli Excel





Work in progress



- Riorganizzazione del database in modo che sia interfacciabile al software di data mining
METABASE

Layout di pagina Formule Dati Revisione Visualizza Guida Team Cosa vuoi fare?														
Booleman Old 11 A A Testo a capo														
Carattere Allineamento Numeri Formattazione Formatta come condizionale Stili Inserisci														
Data														
R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF
HE 38m	PS	22.3	757.2	1.003	2.884				2.594	0.992	1.000	88.80	140.73	140.2
HE 38m	PS	22.3	757.2	1.003	1.525				2.525	0.992	1.000	88.80	148.43	149.8
HE 38m	PS	22.3	757.2	1.003	2.144				2.545	0.992	1.000	88.80	137.83	137.7
HE 38m	PS	22.3	757.2	1.003	0.952				0.953	0.992	1.000	88.80	86.50	87.4
HE 38m	PS	22.3	757.2	1.003	1.392				1.392	0.992	1.000	88.80	81.83	80.7
HE 38m	PS	22.3	757.2	1.003	1.875				2.615	0.992	1.000	88.80	137.25	137.6
HE 38m	PS	22.3	757.2	1.003					4207/02	0.992	1.000	88.80	4207/02	90.2
HE 38m	PS	22.3	757.2	1.003					4207/02	0.992	1.000	88.80	4207/02	90.4
HE 38m	PS	22.3	757.2	1.003	8.04				6.041	0.992	1.000	88.80	472.68	473.2
HE 38m	PS	22.3	757.2	1.003	4.37				4.370	0.992	1.000	88.80	256.89	256.9
HE 38m	PS	22.3	757.2	1.003	1.676				3.676	0.992	1.000	88.80	216.09	216.1
HE 38m	PS	22.3	757.2	1.003	2.696				7.696	0.992	1.000	88.80	432.40	432.2
HE 38m	PS	22.3	757.2	1.003	3.707				3.707	0.992	1.000	88.80	217.91	217.9
HE 38m	PS	22.3	757.2	1.003	1.989				3.989	0.992	1.000	88.80	234.49	234.5





Risultati ottenuti :



- Raccolta e riorganizzazione dati in un unico foglio Excel
 - Progettazione diagramma a stella
 - Creazione database di test in MySQL



Step futuri:



- Caricamento nuovo database su Metabase
 - Creazione query di interesse





GRAZIE !!!

Vitaliano Ciulli
Pier Giulio Lenzi
Andrea Ceccarelli
Andrea Cioni
Maria Vittoria Garzelli



Fisica Medica della Radioterapia di Careggi

