Deprojecting The AGN Universe
within the Virtual Observatory

Omar Laurino

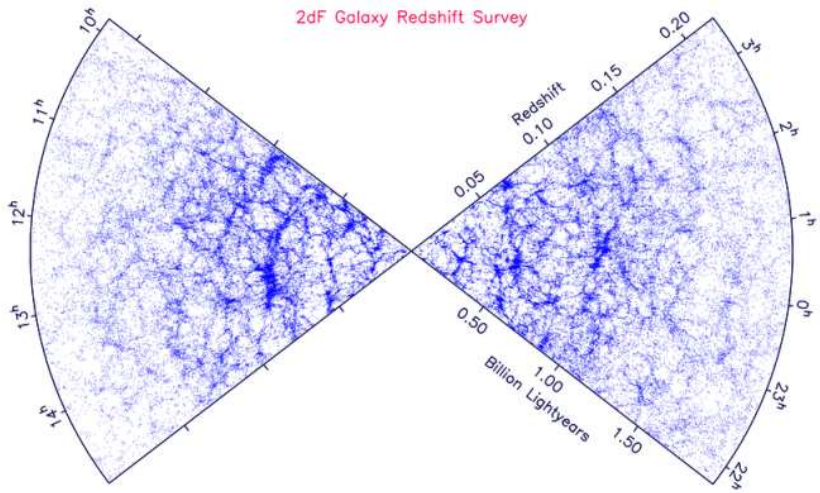
Relatori
Prof. G. Longo
Dott. R. D'Abrusco

Napoli 17 Giugno 2009

Outline

La Struttura dell'Universo in 3D

L'investigazione della struttura su larga scala dell'universo richiede

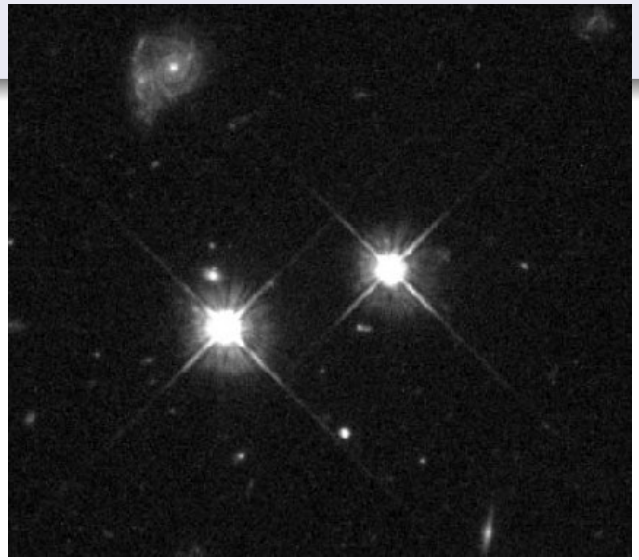


1. Individuazione dei traccianti
   $\longrightarrow$ Quasar
2. Misurazione della distanza
   $\longrightarrow$ Redshift (fotometrici)
3. Statistiche su quantità globali
   $\longrightarrow$ Funzione di luminosità

Traccianti

I quasar

- Oggetti "puntiformi" (*quasi stellari*).
- Spettri non termici.
- Strutture estese se osservate nelle frequenze radio.
- Redshift $z = 0.1 \div 6$.
- Luminosità fino a $10^{14} L_{\odot}$.



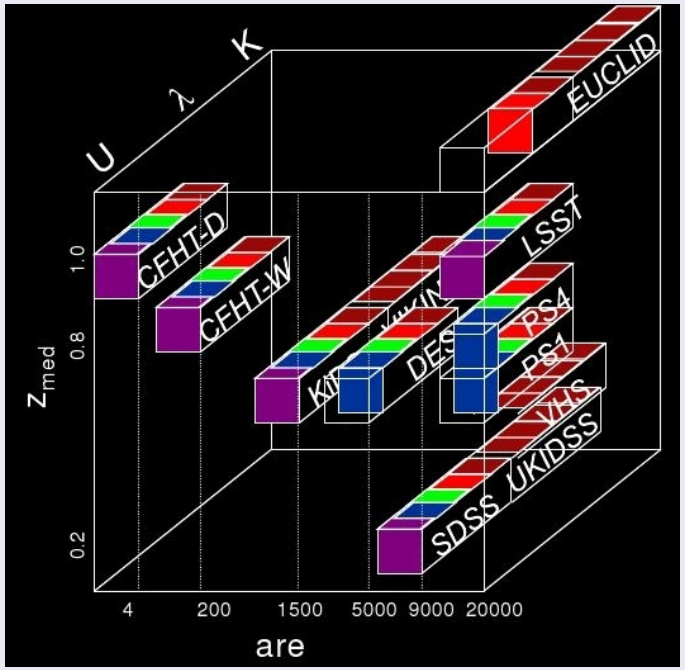
Survey Digitali e Massive Datasets

Le survey digitali a larga scala sono ormai la fonte principale di informazione astronomica: $>$ 100 TB/anno, in rapida crescita.

Ordini di grandezza

- Genoma umano 1 TB
- Congresso USA 20 TB
- Second Life 34 TB
- HathiTrust 78 TB
- Internet Archive 3 PB
- LHC 15 PB
- Conoscenza umana 50 PB
  (Kevin Kelly, The New York Times)

Lo spazio dei parametri



Virtual Observatory: il volto nuovo dell'Astronomia

Osservazioni mirate ed eterogenee
MB-GB

Pochi oggetti
$10^1 - 10^3$

Survey estese omogenee
100TB

Molti oggetti
$10^6 - 10^9$

Cataloghi TB

Molte survey estese e omogenee
$100 - 1000$TB

Molti oggetti in comune a più survey

Cross Matching di cataloghi

Data Mining in Astronomia

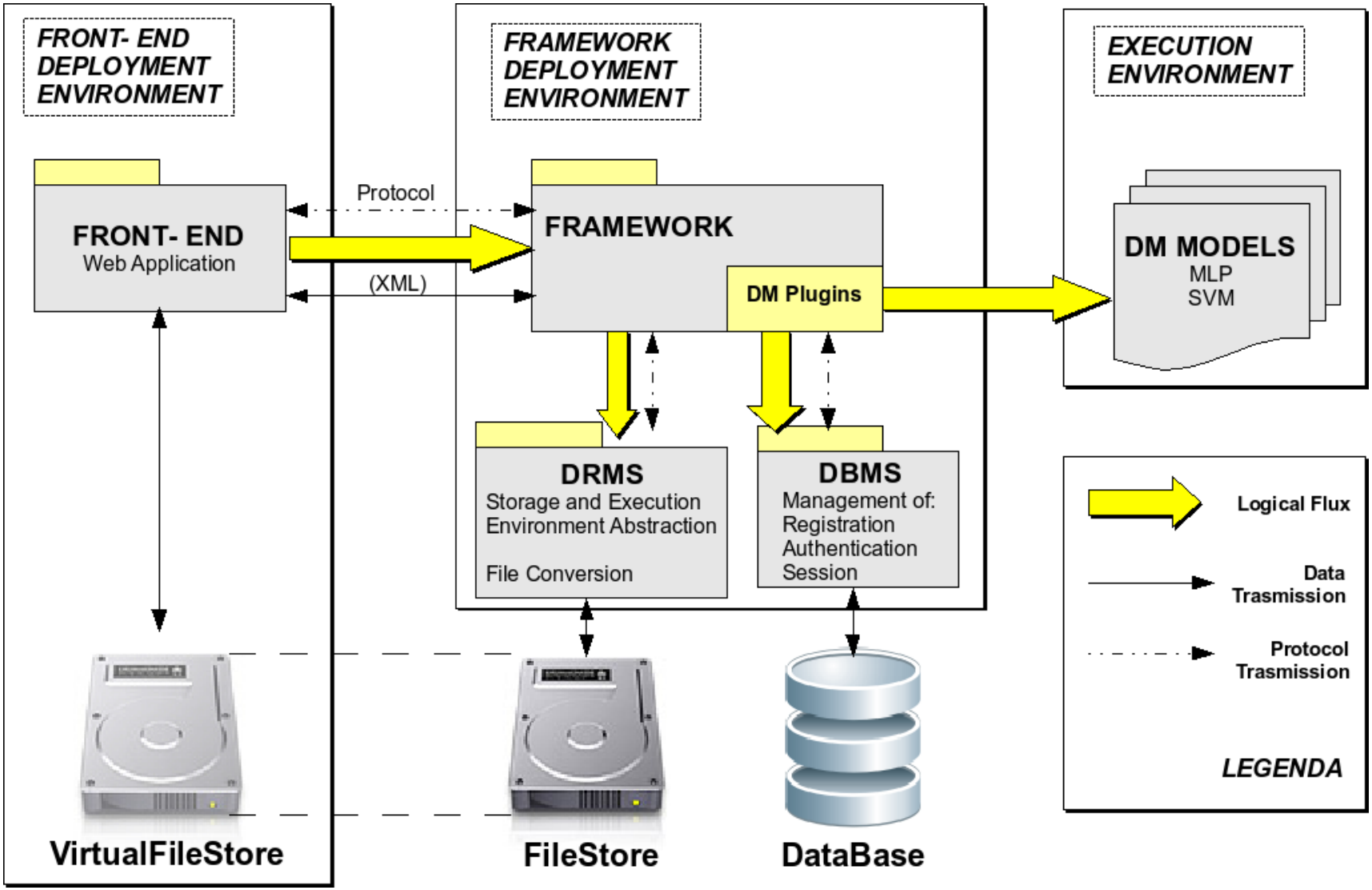
Gli astronomi hanno sempre fatto Data Mining

- Catalogare il conosciuto (classificazione)
- Caratterizzare lo sconosciuto (clustering)
- Trovare relazioni funzionali (regressione)
- Trovare eccezioni (*outliers*)

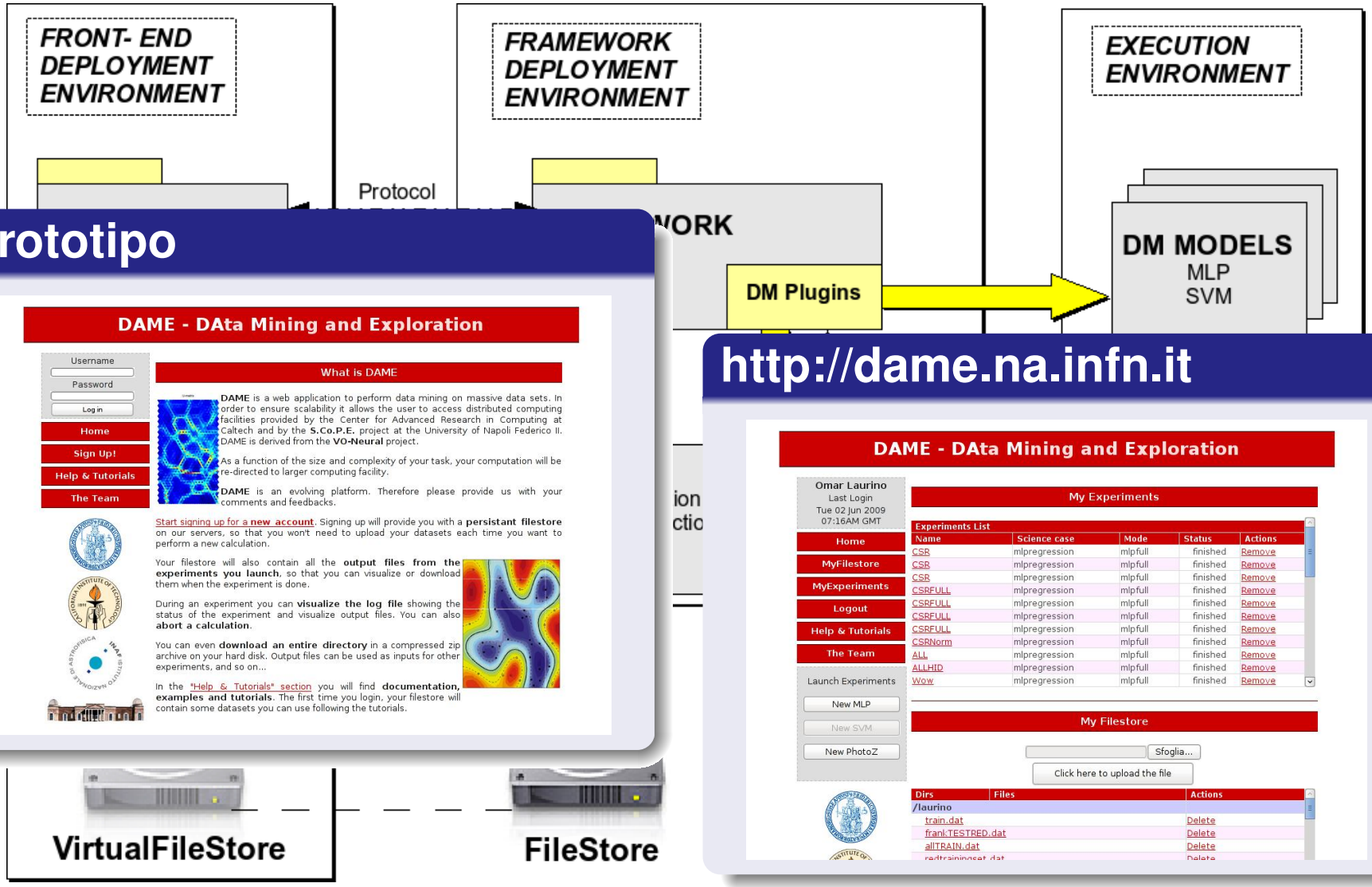
Osservazione → comprensione

1. Osservazione (dati grezzi)
2. Misura parametri (catalogo)
3. Data mining (conoscenza)
4. Comprensione fisica

Data Mining and Exploration (DaME) aka VONeural



Data Mining and Exploration (DaME) aka VONeural



Prototipo

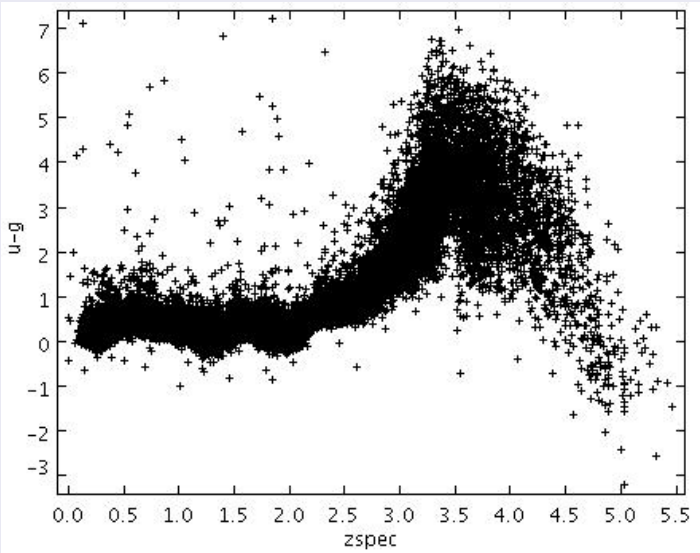
http://dame.na.infn.it

Misurazione delle distanze

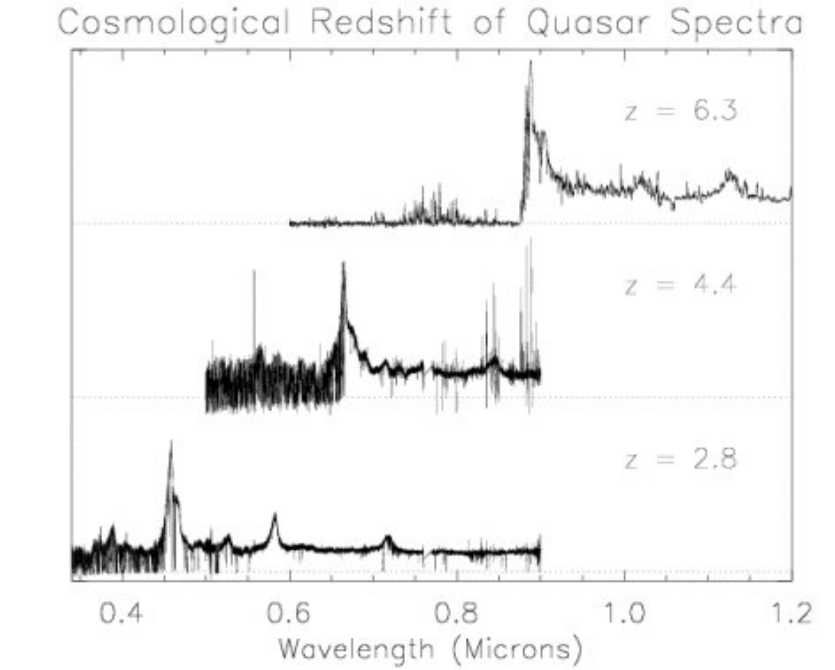
Legge di Hubble

$$z \equiv \frac{\Delta\lambda}{\lambda_0} = \frac{a_0}{a_t} - 1 \approx \frac{v}{c} = \frac{H_0 d}{c}$$

Redshift Fotometrico

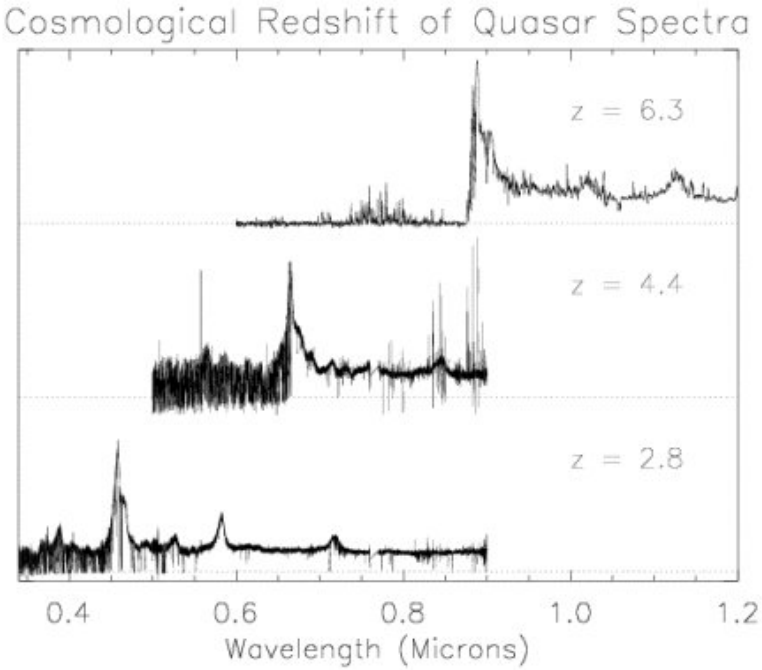


Redshift



z_{spec}: poche sorgenti, maggiore accuratezza. z_{phot}: molte sorgenti, minore accuratezza.

Redshift fotometrici: caratterizzazione fisica



Lo spettro dei quasar

- Continuo $\rightarrow$ legge di potenza
- Lyman α forest
- Redshift fra 0.1 e 6
- Righe di emissione
- Appiattimento

Lo spettro delle galassie

È essenzialmente termico, gli indici di colore sono ben caratterizzati, l'estensione in redshift è un ordine di grandezza inferiore.

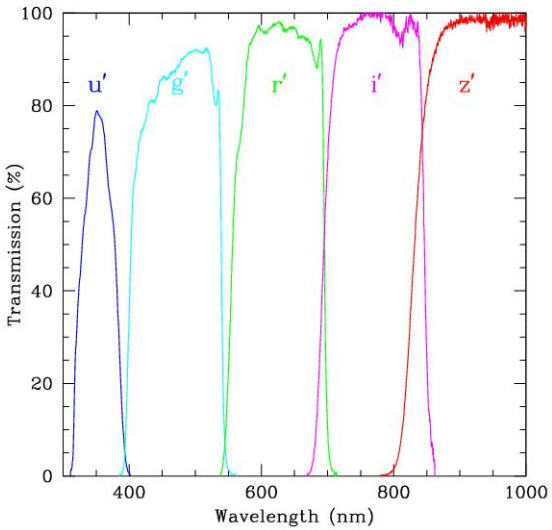
La relazione colore-redshift

$$m_x = -2.5 \log F_x + C$$

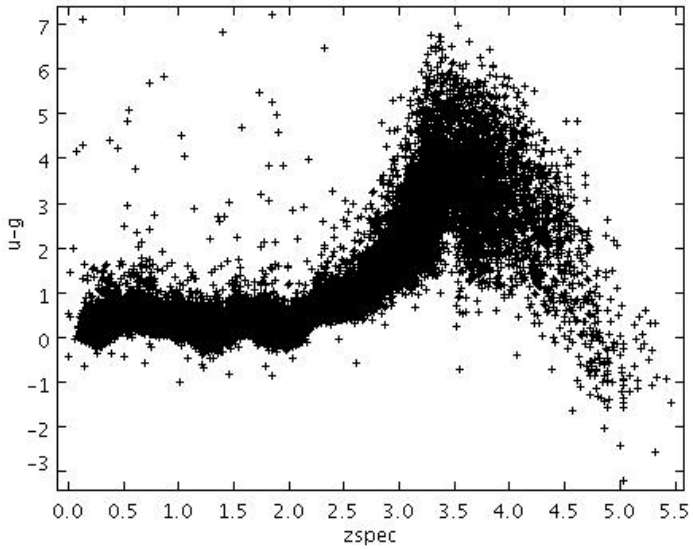
$$c_{xy} = m_y - m_x = -2.5 \log \frac{F_y}{F_x}$$

Le caratteristiche spettroscopiche peculiari dei quasar, integrate nell'informazione fotometrica, possono essere ricostruite tramite gli indici di colore, ovvero tramite i rapporti fra flussi in bande differenti.

Filtri SDSS

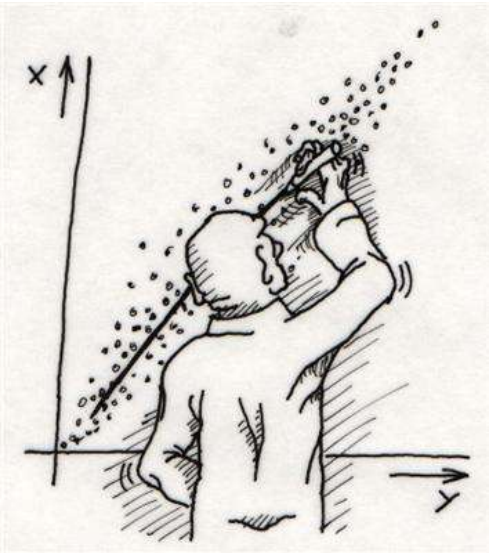


Oltre ad essere "larghe" le bande non sono perfettamente separate.



La relazione fra il singolo colore e il redshift non è iniettiva ed è molto rumorosa.

Percettrone Multistrato (MLP)



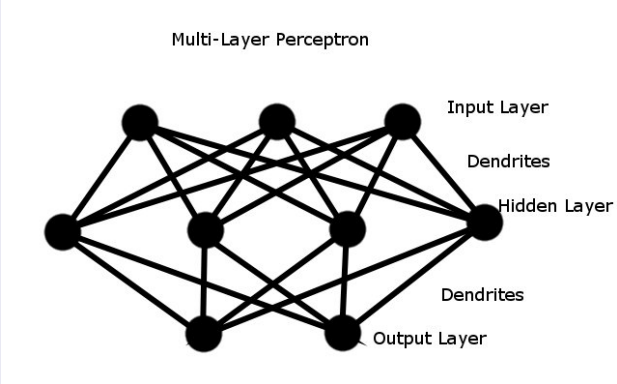
Regressione

La regressione consiste nel predire il valore della variabile dipendente $\mathbf{d} \in \mathbb{R}^N$ a partire da un vettore di ingresso $\mathbf{x} \in \mathbb{R}^M$ composto di M variabili casuali, campionate da una specifica funzione di probabilità.

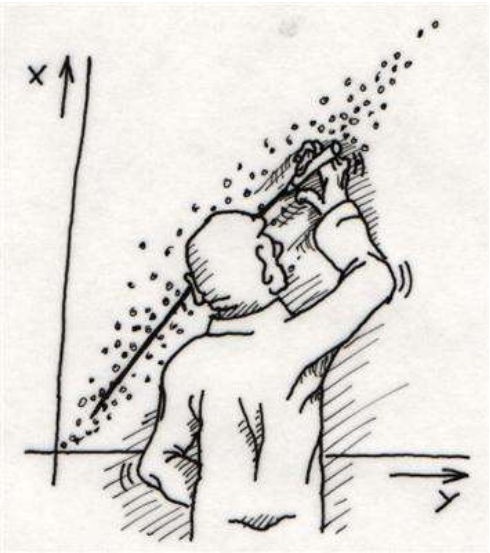
Reti neurali

Una rete neurale può rappresentare un'arbitraria forma funzionale come sovrapposizione di sigmoidi. Specifici algoritmi di addestramento consentono di individuare il modello che meglio interpola i dati, garantendo un buon grado di generalizzazione.

MLP



Percettrone Multistrato (MLP)



Regressione

La regressione consiste nel predire il valore della variabile dipendente $\mathbf{d} \in \mathbb{R}^N$ a partire da un vettore di ingresso $\mathbf{x} \in \mathbb{R}^M$ composto di M variabili casuali, campionate da una specifica funzione di probabilità.

Reti neurali

Una rete neurale può rappresentare un'arbitraria forma funzionale come sovrapposizione di sigmoidi. Specifici algoritmi di addestramento consentono di individuare il modello che meglio interpola i dati, garantendo un buon grado di generalizzazione.

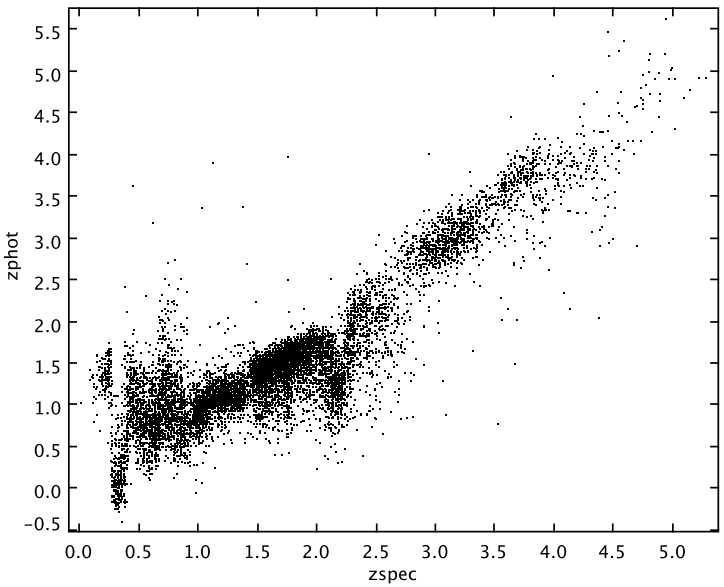
Fasi di addestramento

- Training
- Validation
- Test

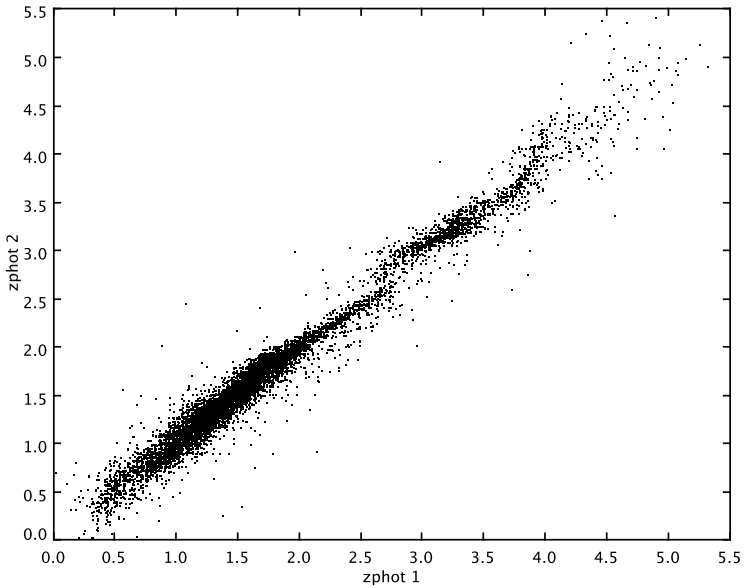
Problemi nell'uso di una rete singola

- Compromesso varianza/bias
  $\longrightarrow$ Committee
- Diverse funzioni in diversi sottodomini
- Diversi regimi di errore in diversi sottodomini
  $\longrightarrow$ Mixture of experts

Ricostruzione z_{phot} con singola rete



Confronto ricostruzioni da singole reti

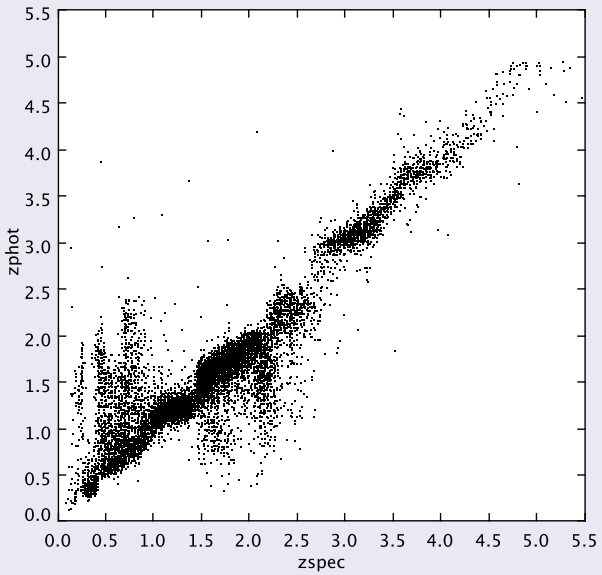


Weak Gated Experts

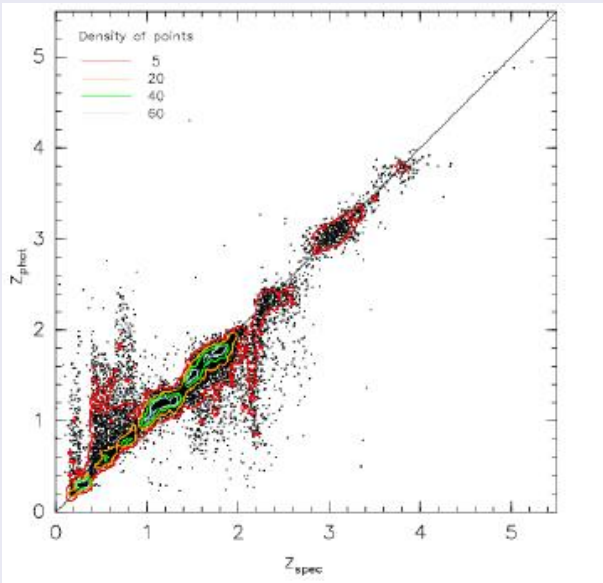
- Reti diverse possono essere addestrate su diversi sottodomini dello spazio dei parametri (*experts*).
- Un'ulteriore rete viene addestrata a combinare opportunamente l'uscita dei singoli esperti (*gating network*).
- L'accuratezza e la capacità di generalizzazione di una rete di esperti non è inferiore a quelle della rete "migliore", punto per punto nello spazio dei parametri.
- Il partizionamento avviene per mezzo di un algoritmo non supervisionato di *clustering fuzzy*.

Quasar – Dati ottici

Questa Tesi

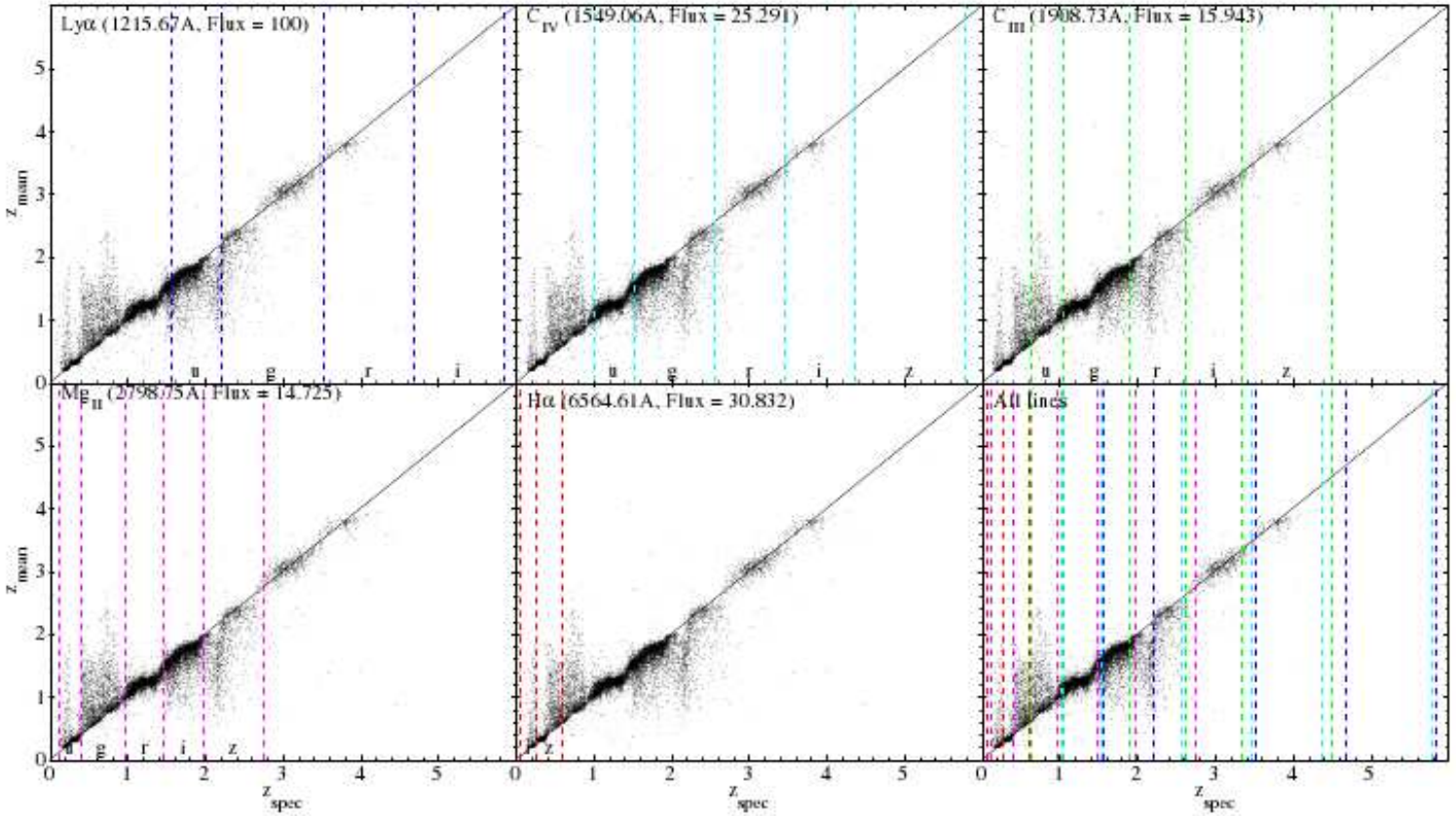


Ball 2008 – *k*NN (con PDF)



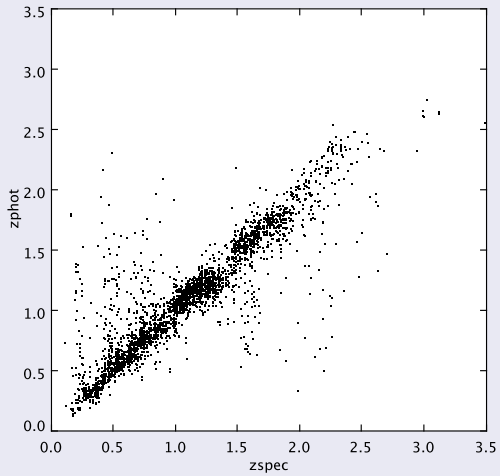
Il nostro metodo è confrontabile col *k*NN, sebbene sia estremamente più efficiente. Il *k*NN-PDF processa quattro sorgenti al secondo su 100 processori!!!

L'origine delle degenerazioni

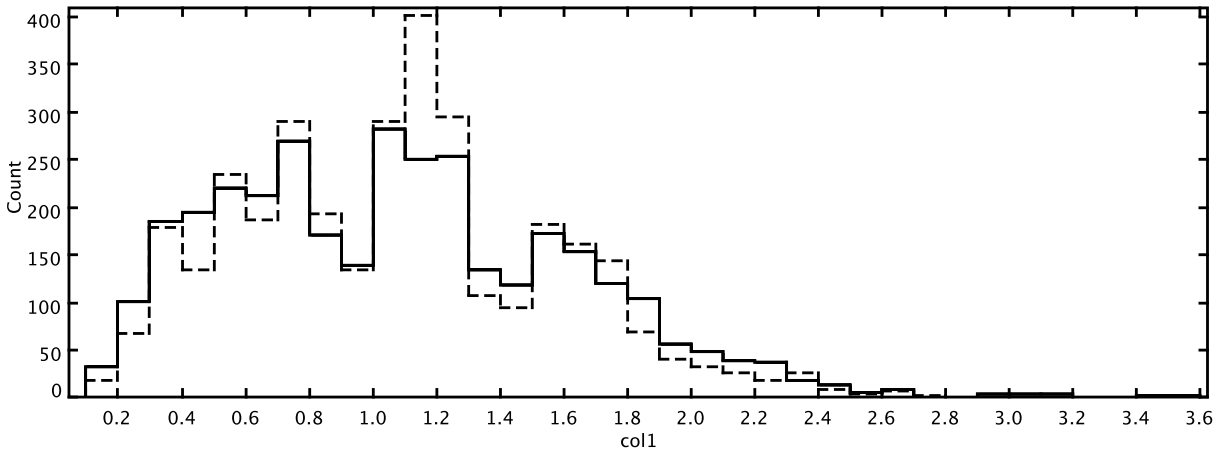


È evidente l'effetto del passaggio delle *feature* spettroscopiche tra una banda e l'altra del sistema fotometrico. Ancora più determinante l'uscita (o l'ingresso) di determinate *feature* dal (nel) sistema fotometrico.

Quasar – Dati ottici e ultravioletti



Estendere lo spazio dei parametri consente di risolvere molte degenerazioni, migliorando sensibilmente la ricostruzione del redshift. In particolare le osservazioni UV (GALEX) consentono di misurare la Lyman-α forest. Per $z > 2.5$ il numero di QSO osservati è esiguo.

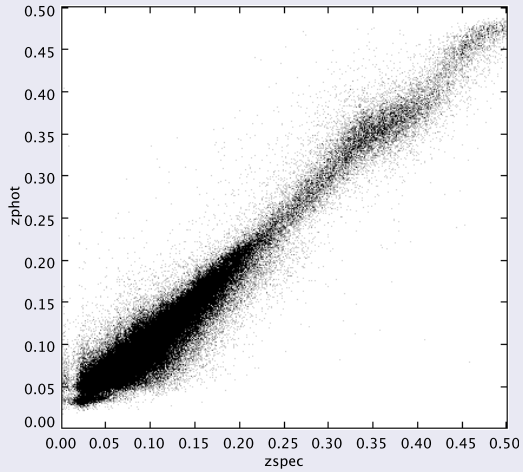


Distribuzione redshift.

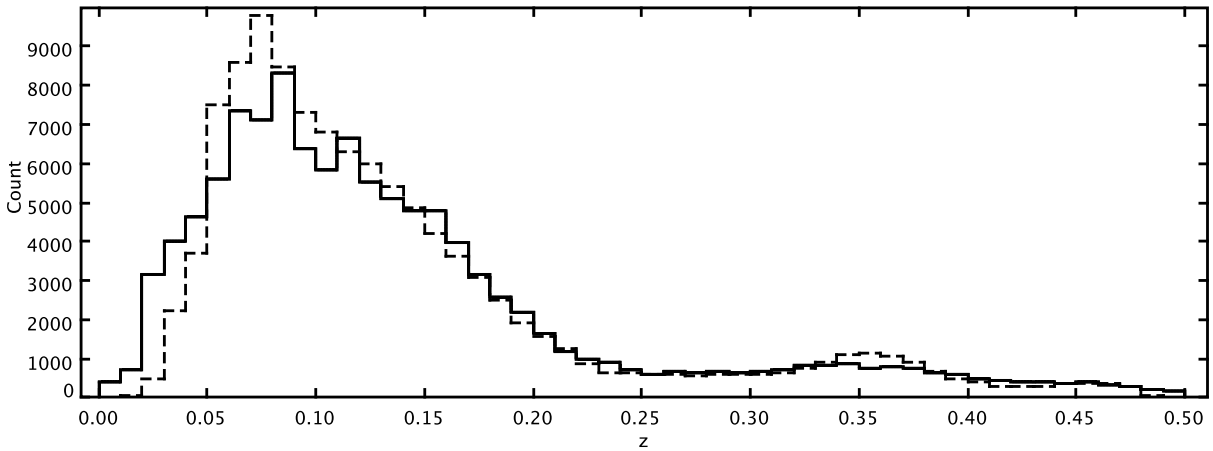
z_{phot}: linea tratteggiata.

z_{spec}: linea solida.

Galassie – Dati ottici



Le galassie emettono perlopiù nell'ottico, hanno spettri termici e il loro redshift è inferiore rispetto ai quasar. L'applicazione dei WGE sulle galassie migliora i risultati in letteratura. Una procedura di interpolazione può essere impiegata per ridurre effetti di *bias*.



Distribuzione redshift.

z_{phot}: linea tratteggiata.

z_{spec}: linea solida.

Tabella riassuntiva

Tabella comparativa degli algoritmi: *k*NN (Ball, 2008), CZR (Weinstein, 2004), WGE (questo lavoro)
S – Dati ottici (SDSS)
G – Dati UV (GALEX)

Method	Dataset	Variance	$\frac{\sigma^2}{1+z}$	$\mu\left(\frac{\Delta z}{1+z}\right)$	$\%\Delta_{0.1}$	$\%\Delta_{0.2}$	$\%\Delta_{0.3}$
*k*NN	S	**0.123**	**0.034**	0.095	54.9	73.3	80.7
*k*NNPDF	S	–	–	–	53.8	72.4	79.8
CZR	S	0.265	0.079	0.115	**63.9**	**80.2**	**85.7**
WGE	S	0.142	0.059	0.032	48.8	70.3	78.9
WGE+err	S	0.133	0.056	**0.025**	48.7	71.4	80.4
*k*NN	SG	**0.054**	**0.014**	0.060	70.8	85.8	90.8
*k*NNPDF	SG	–	–	–	71.8	86.4	90.8
CZR	SG	0.136	0.031	0.071	**74.9**	**86.9**	91.0
WGE	SG	0.058	0.030	0.022	67.9	85.2	91.1
WGE+err	SG	0.057	0.029	**0.012**	69.3	86.2	**91.3**

La funzione di luminosità dei quasar

La FL esprime la densità di sorgenti in funzione del redshift (se c'è evoluzione) ed in funzione della magnitudine assoluta. La forma della FL fornisce informazioni sui processi evolutivi dei quasar (es. pura luminosità o pura densità).

FL da volumi accessibili (binning)

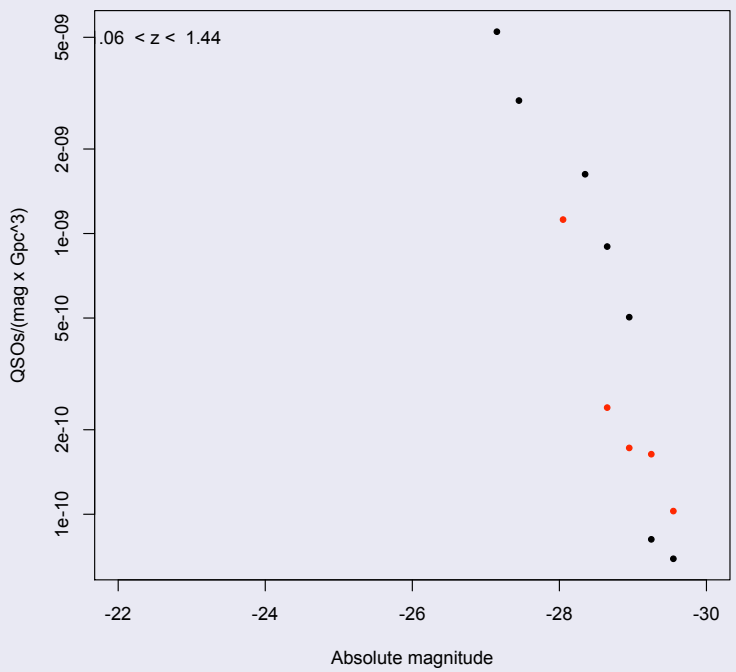
$$\Phi(M, z) = \frac{1}{dM} \sum_{j=1}^{n} \frac{1}{V_{m,j}}.$$

Ciascun quasar contribuisce alla densità della popolazione tramite l'inverso del suo "volume accessibile": in ciascun bin di redshift un quasar di magnitudine assoluta M ha "accesso" ad un guscio $V_m = V(z_1, z_2)$.

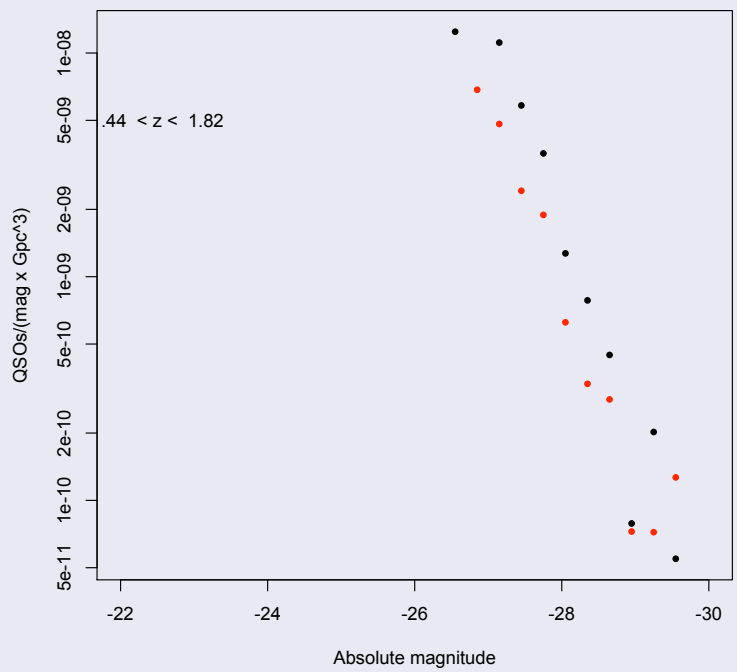
La funzione di luminosità dei quasar

Confronto fra la FL calcolata rispettivamente con i redshift spettroscopici e con i redshift fotometrici, per valutare il peggioramento introdotto dall'uso di questi ultimi. In rosso i punti con z_{phot}, in nero quelli con z_{spec}.

$1.06 < z < 1.44$



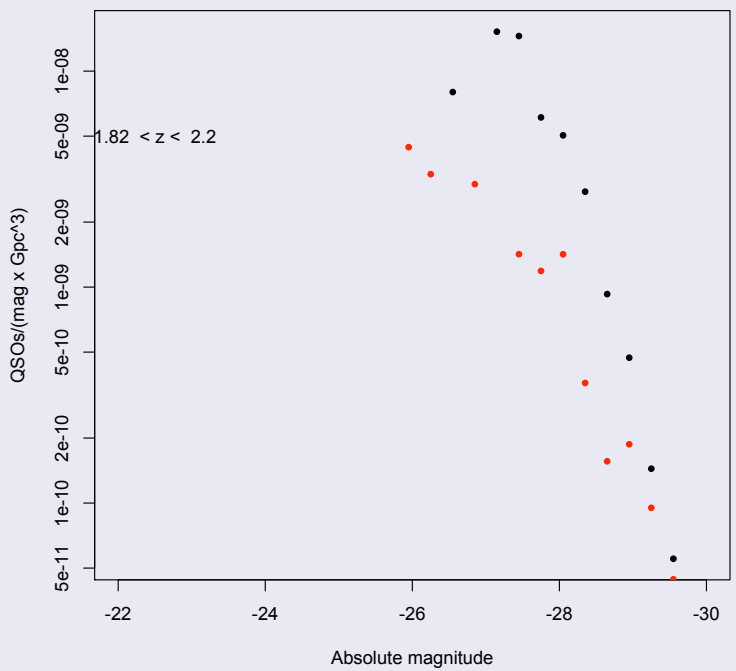
$1.44 < z < 1.82$



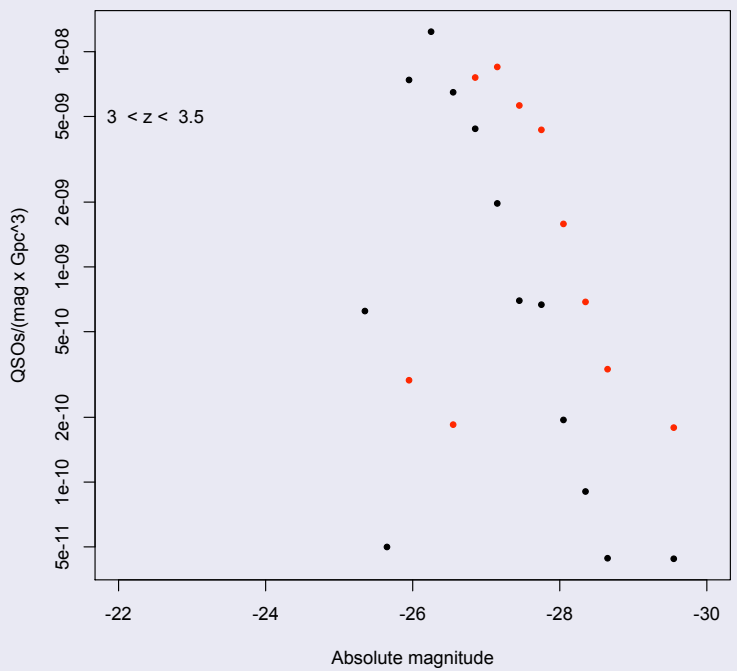
La funzione di luminosità dei quasar

Confronto fra la FL calcolata rispettivamente con i redshift spettroscopici e con i redshift fotometrici, per valutare il peggioramento introdotto dall'uso di questi ultimi. In rosso i punti con z_{phot}, in nero quelli con z_{spec}.

$1.82 < z < 2.20$



$3.0 < z < 3.5$



Barre di errore

Calcolare le barre di errore sulle predizioni delle reti neurali è un processo delicato e complesso. Si distinguono quattro diverse sorgenti di errore:

Input noise. Paragonabile alla normale propagazione degli errori.

Model variance. Diversi modelli producono diverse predizioni (dai dati e dall'addestramento).

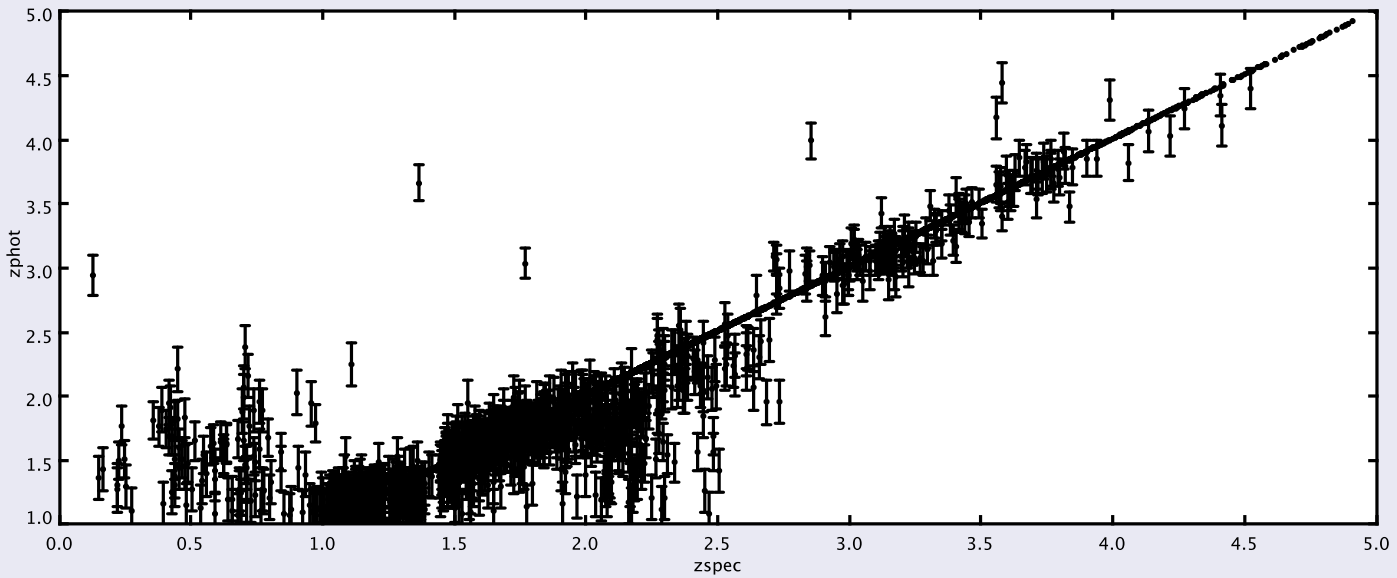
Model bias. Ciascun modello può presentare un certo *bias* (dai dati e dall'addestramento).

Target noise. In alcune regioni dello spazio dei parametri i dati possono rappresentare "male" la relazione funzionale fra *input* e *target*.

Stima dell'errore *a posteriori*

Per ciascun intervallo di redshift (fotometrico) si può stimare l'errore complessivo prodotto dalla ricostruzione sfruttando la base di conoscenza.

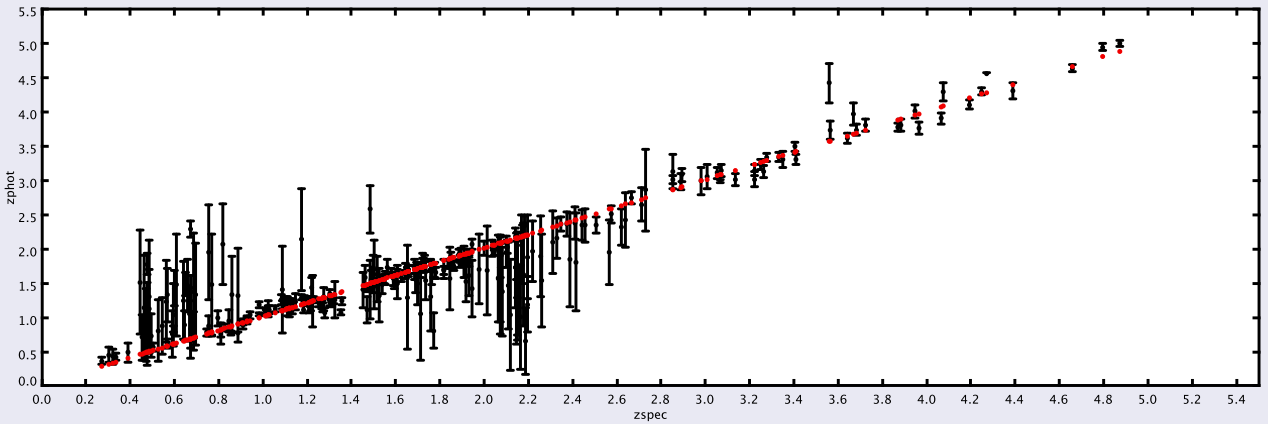
Barre di errore *a posteriori*



Stima del *target error*

Si addestra una rete neurale a ricostruire i residui sulla base degli input (usando il metodo WGE). L'addestramento non è banale a causa delle degenerazioni. Nelle regioni "patologiche" la stima dell'errore può essere paragonata ad un errore massimo.

Barre di *target error*



Conclusioni

Risultati della tesi

- sviluppo di una piattaforma innovativa per il Data Mining (DaME);
- implementazione di un prototipo funzionante di DaME;
- progettazione e implementazione di un algoritmo originale per la stima dei redshift fotometrici dei quasar;
- stima rigorosa delle barre d'errore;
- funzione di luminosità.

Conclusioni

Sviluppi futuri

- Catalogo di redshift fotometrici di $\sim 10^6$ candidati quasar (D'Abrusco et al. 2009)
- Derivazione della funzione di luminosità dei quasar
- Aggiornamento del catalogo di redshift fotometrici delle galassie in D'Abrusco et al. 2007
- Esplorazione di regioni più problematiche dello spazio dei parametri; ad esempio, galassie oltre il limite di completezza spettroscopica della SDSS.

L'uso di metodi quali il WGE consente di incrementare di almeno un ordine di grandezza gli oggetti utilizzati per derivare statistiche globali (*candidati*), e ciò è di notevole importanza per la cosmologia osservativa.

Ricerca sul campo