RUHR-UNIVERSITÄT BOCHUM

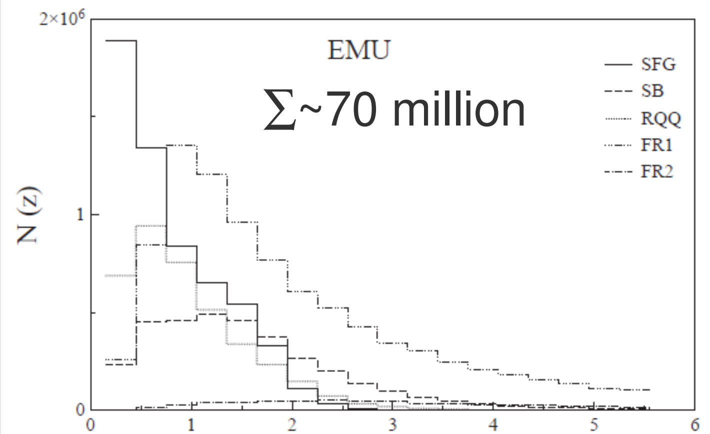
Bringing order to chaos – Machine-learning in astronomy

Peter-Christian Zinn
& Kai Lars Polsterer, Fabian Gieseke
& Enno Middelberg, Ray Norris,
Ralf-Jürgen Dettmar

Why do we need new data handling techniques?

1: New radio surveys will produce lots of data!

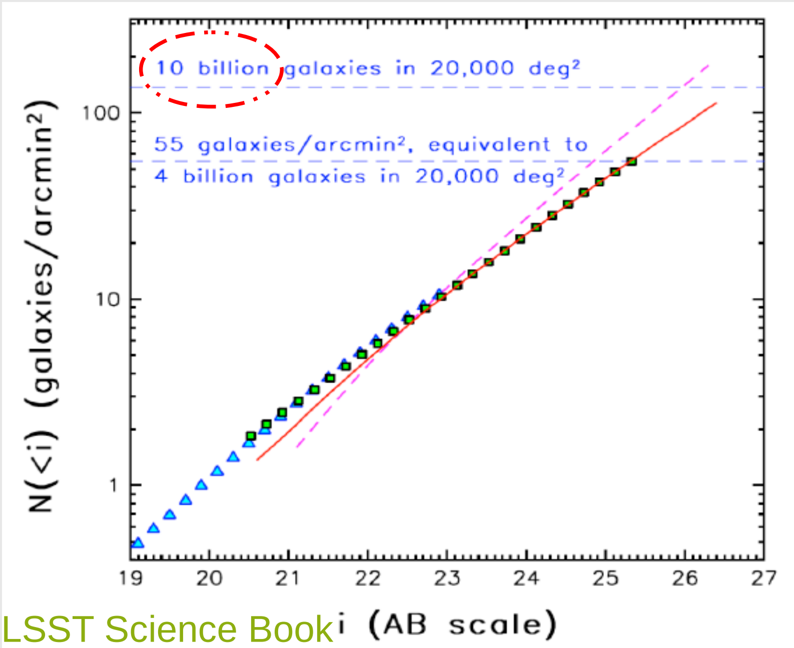
- ASKAP/EMU ~ 70 million sources
- LOFAR/Tier 1 ~ 7 million sources
- WSRT/WODAN ~ 10 million sources



Norris et al. 2011

2: New optical/NIR surveys will produce even more data!

- Pan-STARRS/PS1-3π ~ 5-30 billion sources
- LSST/Galaxy "gold sample" ~ 10 billion galaxies



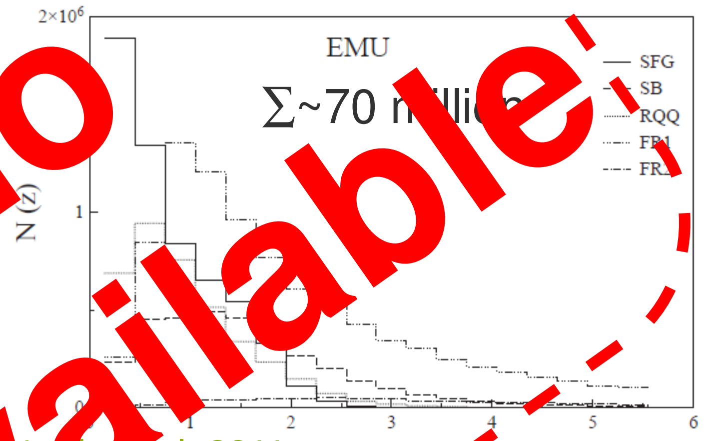
LSST Science Book

Why do we need new data handling techniques?

Mostly no spectra available!

1: New radio surveys will produce lots of data!

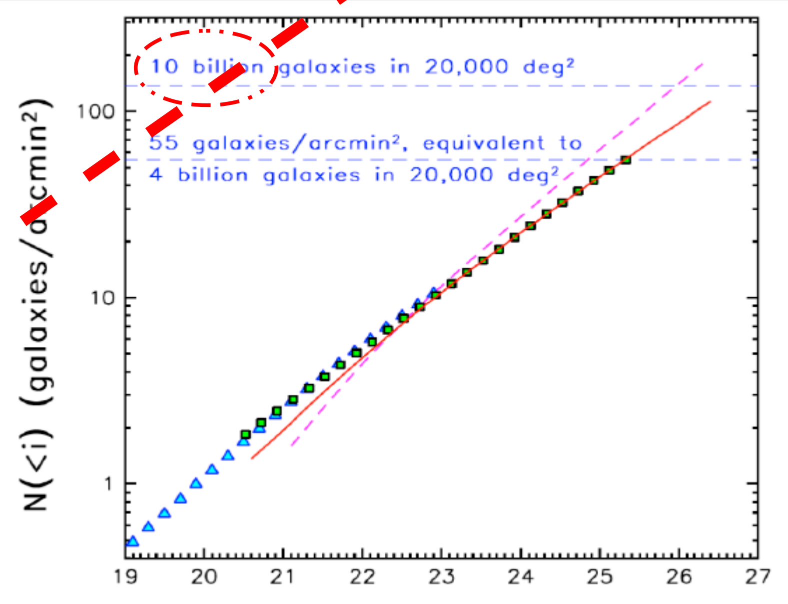
- ASKAP/EMU ~ 70 million sources
- LOFAR/Tier 1 ~ 7 million sources
- WSRT/WODAN ~ 10 million sources



Norris et al. 2011

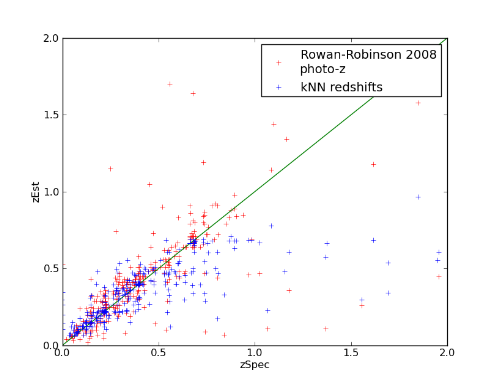
2: New optical/NIR surveys will produce even more data!

- Pan-STARRS/PS1-3π ~ 5-30 billion sources
- LSST/Galaxy "gold sample" ~ 10 billion galaxies



LSST Science Book

Implications for survey science



1: There are no spectroscopic redshifts

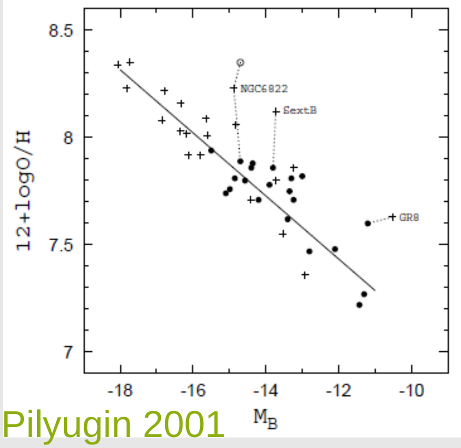
- Redshift information must be accessed on other ways → photometric (better: statistical) redshifts

2: There are no spectral classifications

- Classification of an object must be inferred on other ways → Flux ratios or SED-fitting (better: kNN classification) becomes more important

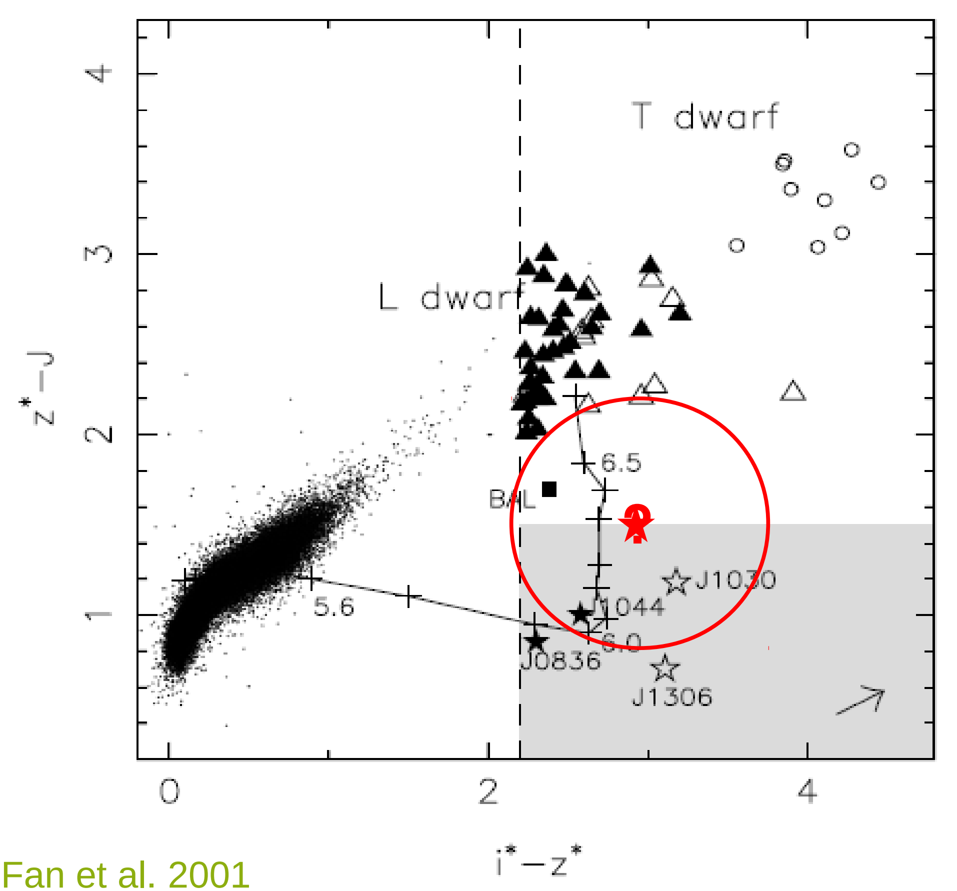
3: There are no spectroscopically derived parameters

- Classic parameters such as metallicity must be derived on other ways → scaling relations (better: kNN regression) must be utilized



Pilyugin 2001

The fundament: the k nearest neighbor algorithm



Fan et al. 2001

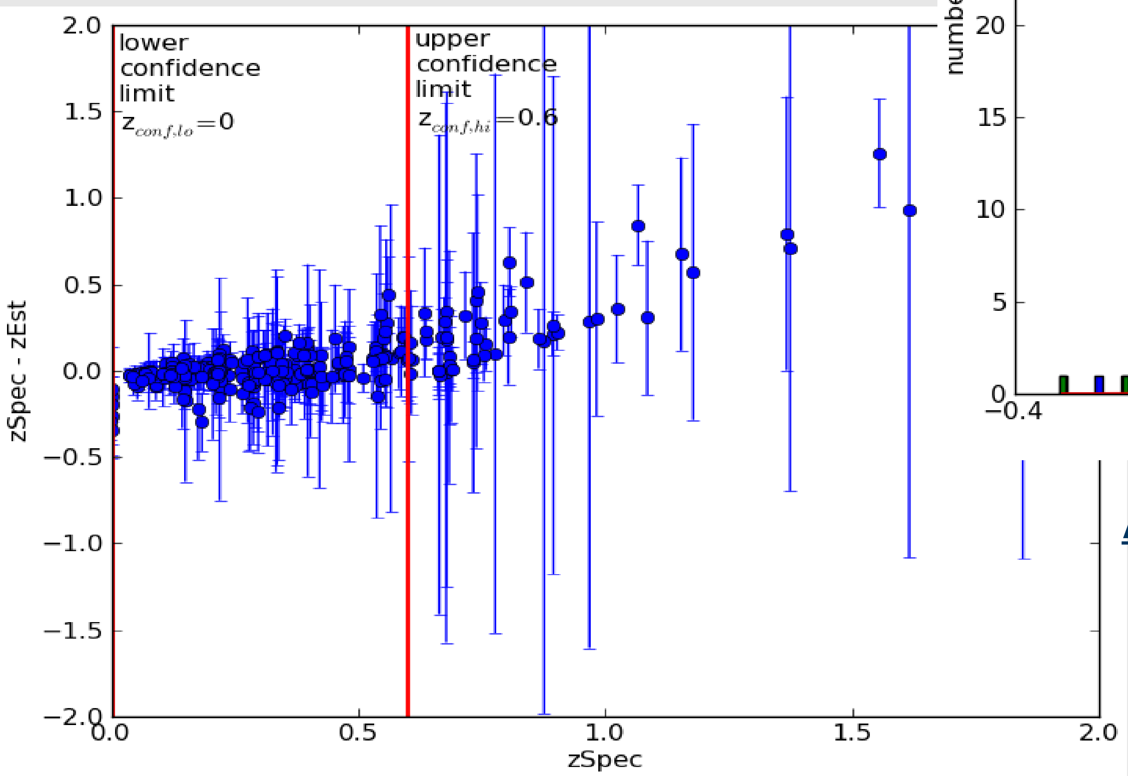
The task: Tune the algorithm parameters such that the optimal selection results!

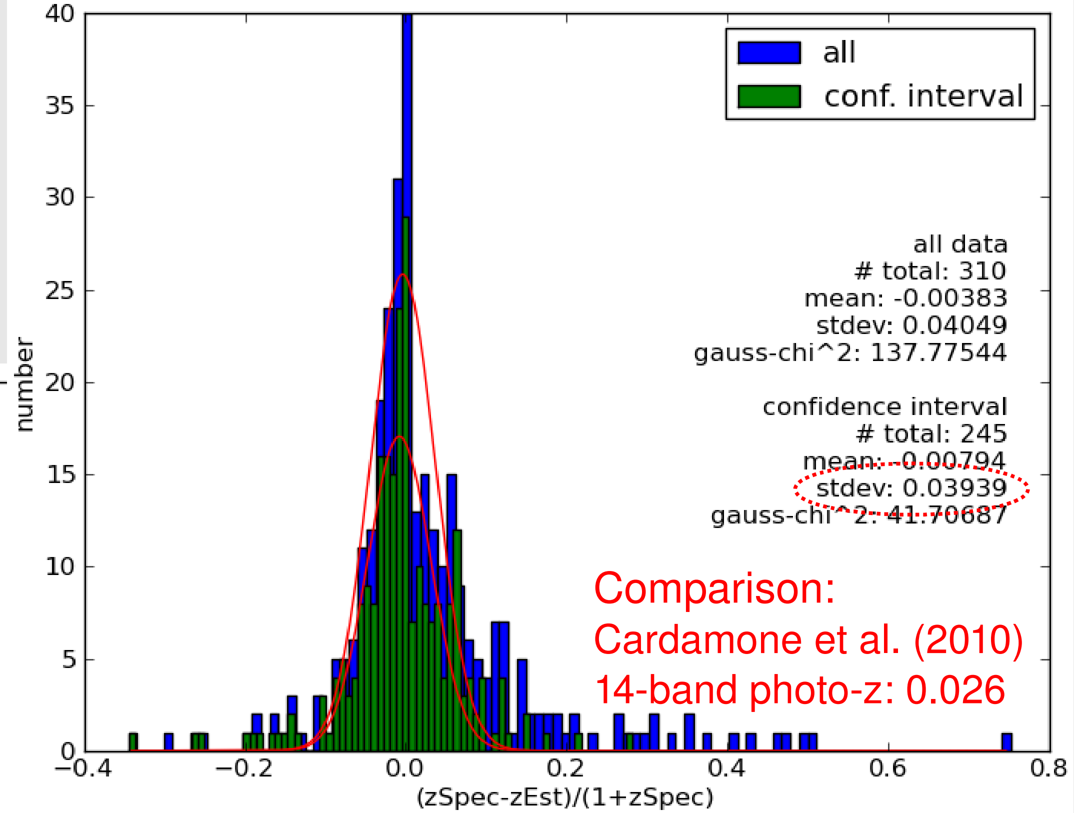
- kNN could be regarded as expansion of traditional color selection criteria
- Example: Fan et al. Selection of high-z quasars in the SDSS
  - Problems:
    + Selection only possible in 2-dimensional feature space
    + Selection criteria must be well-known
    + No information about quality of selection
- Advantages of kNN-based approach:
  - The computer can handle n-dimensional features spaces
  - The algorithm need not know any astrophysical selection criterion
  - Quality measurement and prediction of values (e.g. redshift) and corresp. errors possible

Example 1: statistical redshifts

Statistical redshifts for ATLAS

- ATLAS has spec-z for ~30% of all sources
- Stat-z trained with **12-band data** (ugriz,IRAC,MIPS24,13cm,20cm)

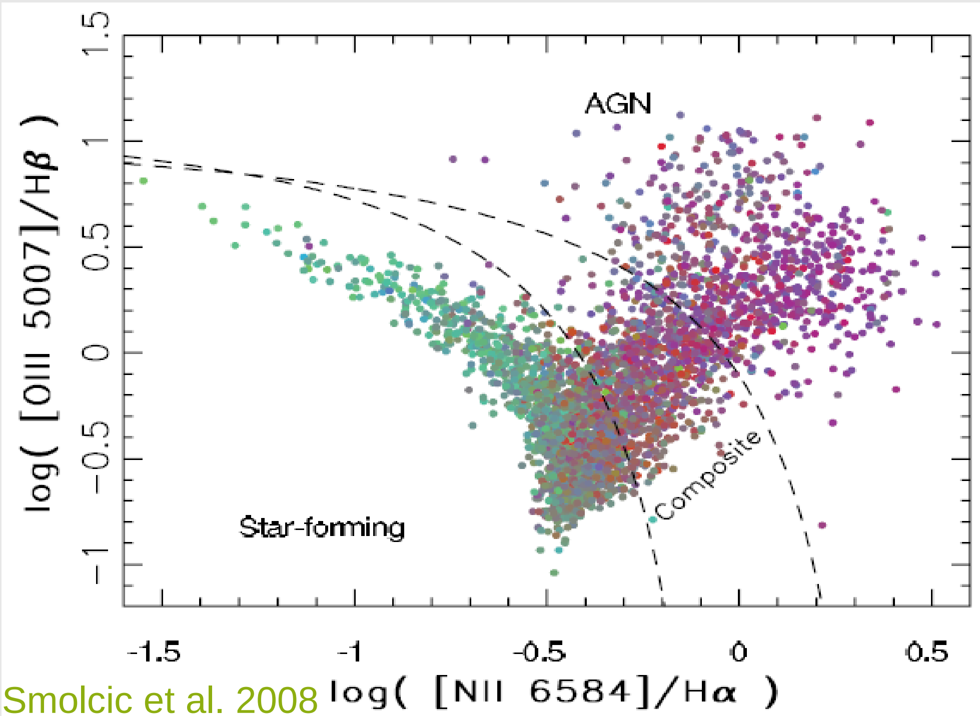




Advantages of statistical redshifts

- **No assumptions** must be made (no template SEDs, luminosity range, ...)
- Computation **much faster** than for classical photo-z ($t_{stat\text{-}z} \sim n*\log(n)$ | $t_{photo\text{-}z} \sim n^{\alpha}$)

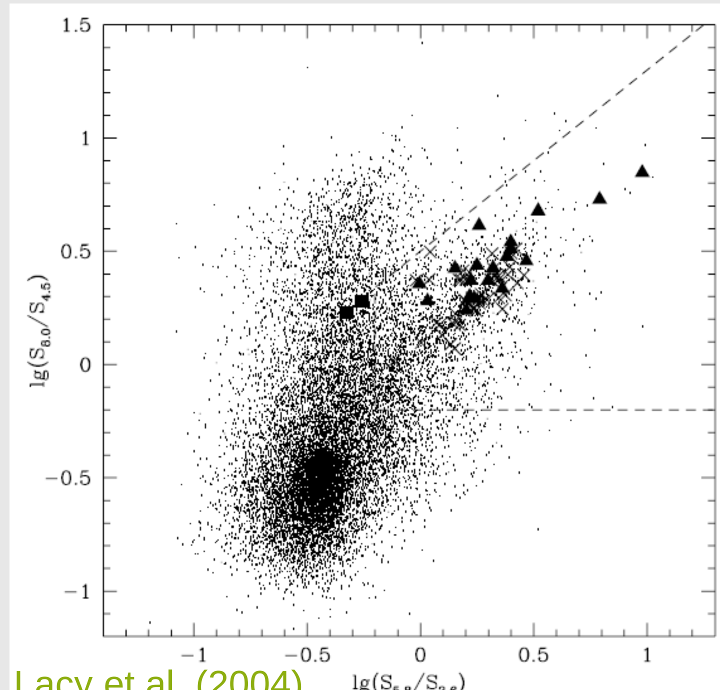
Example 2: Object classification



Smolcic et al. 2008

- Example: star-formation/AGN separation
- Classical tool: **Baldwin-diagram** (requires spectroscopy)
- Alternative: **MIR color-color selection** (not very reliable)

- kNN-based classification of test-sample in the COSMOS field yields combined **false classification rate of 11.1%**
- For comparison, Smolcic et al. (2008) achieve **contamination rates between 15% - 20%** using a highly sophisticated photometric selection method

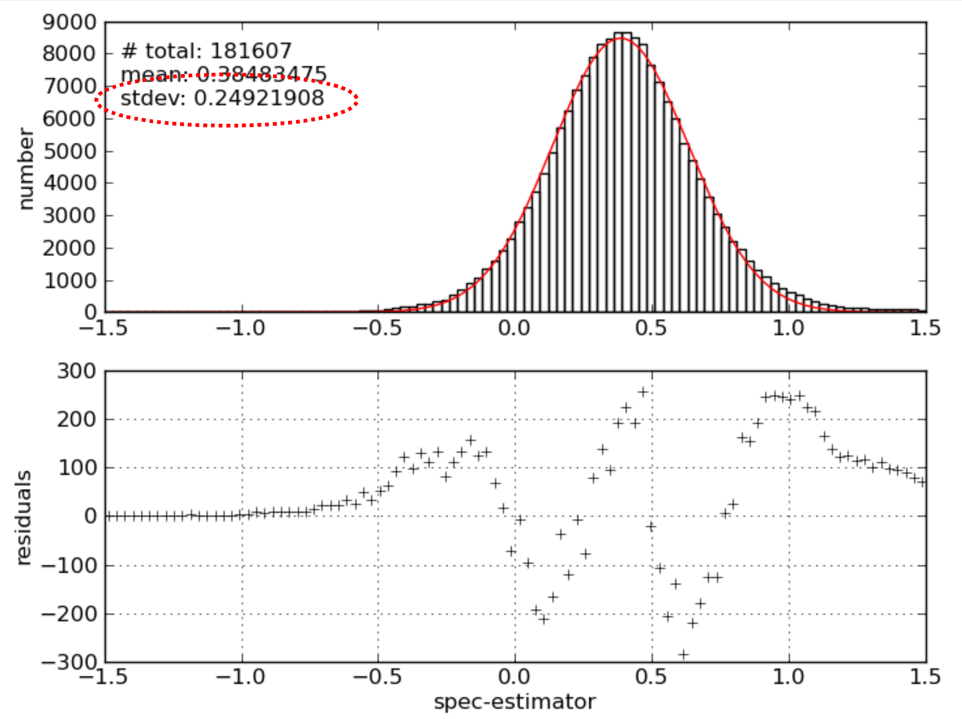


Lacy et al. (2004)

Example 3: metallicity

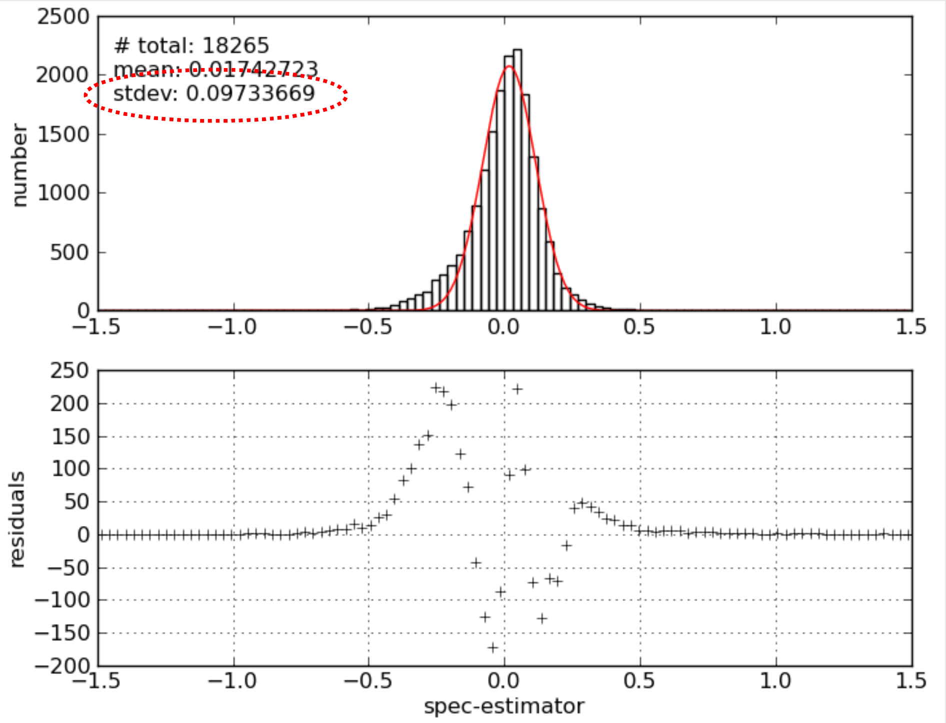
Metallicity from L-Z relation

- Spectroscopic input: SDSS metallicities as derived by Brinchman et al. (2004)
- L_r-Z relation calibrated by the 2dF survey (Lamareille et al. 2004) applied to Galactic extinction-corrected fluxes
- No other assumptions made



Metallicity from kNN regression

- Spectroscopic input: SDSS metallicities as derived by Brinchman et al. (2004)
- kNN regression with respect to the 90 nearest neighbors, no measurement errors taken into account
- No other assumptions made



Summary

- We presented the first results of utilizing **advanced machine-learning techniques** to classify/analyze large data sets.
- Dealing with large data sets will become increasingly important due to the **enormous amounts of data** forthcoming (radio) surveys will produce.
- The **k nearest neighbor-based approach was tested** on available data from ATLAS, COSMOS and the SDSS.
- Results for redshifts, object classifications and the regressional computation of astrophysical quantities (e.g. metallicity) all yield **promising results**.
- kNN-based approach will be used by several surveys, the largest one being **ASKAP/EMU**.