RUHR-UNIVERSITÄT BOCHUM

Bringing Order to Chaos

new data-mining techniques for new surveys

Peter-Christian Zinn

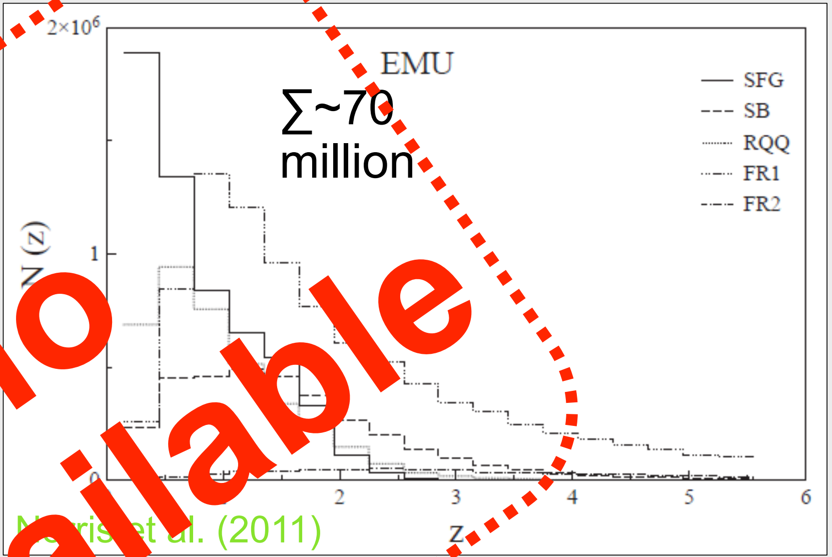
Astronomical Institute of Ruhr-University, Bochum, Germany

CSIRO Astronomy & Space Science, Sydney, Australia

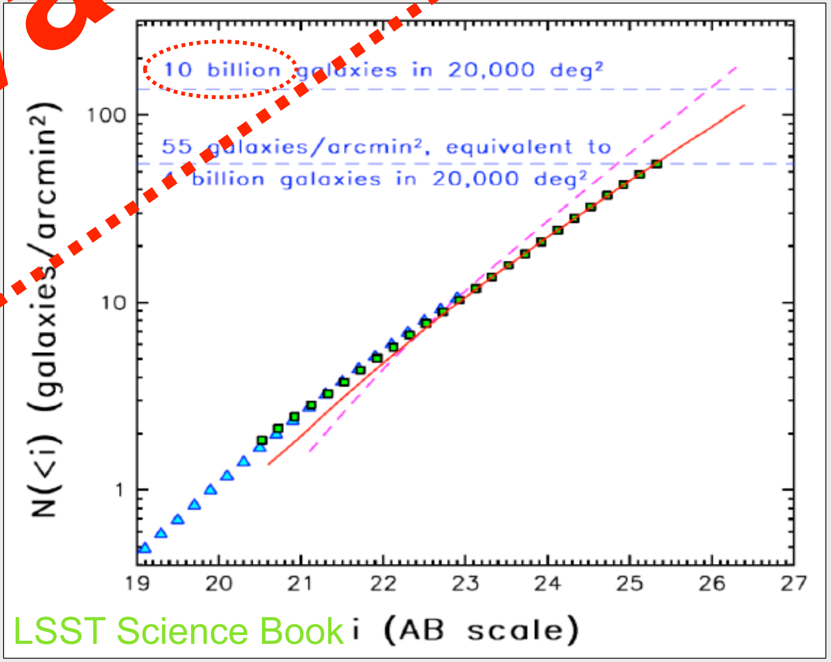
Why new data handling techniques?

- The next generation of photometric surveys will produce lots of data!
  - In 5 yrs: order 1 billion objects
  - In 15 yrs: order 10 billion objects
- New spectroscopic surveys will more than 10-fold the number of spectra compared to present!
  - 4MOST: order 10 million
  - HEXA: order 100 million
- Interesing fact: the ratio doesn't change!
  - Only ~1% of the photometric sources has and will have spectra

Mostly no spectra available



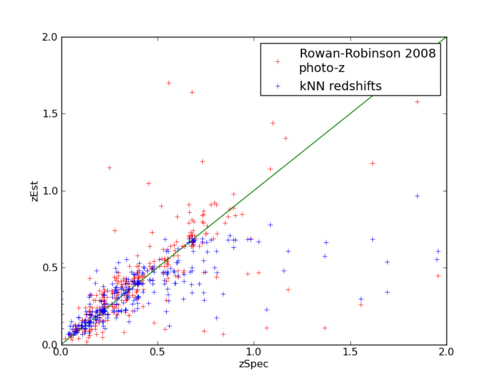
Norris et al. (2011)

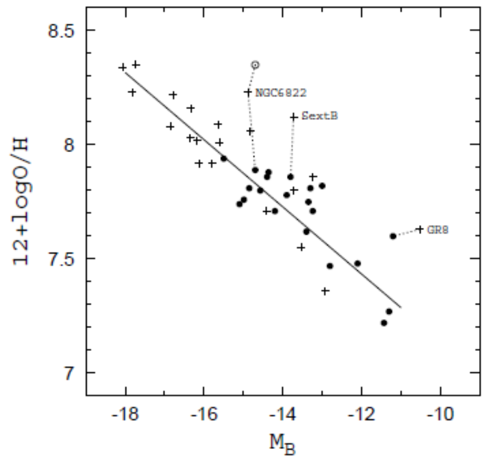


LSST Science Book

Implications for survey science

- There are no spectroscopic redshifts
  - Redshift information must be accessed on other ways → photometric (better: statistical) redshifts
- 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
- 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

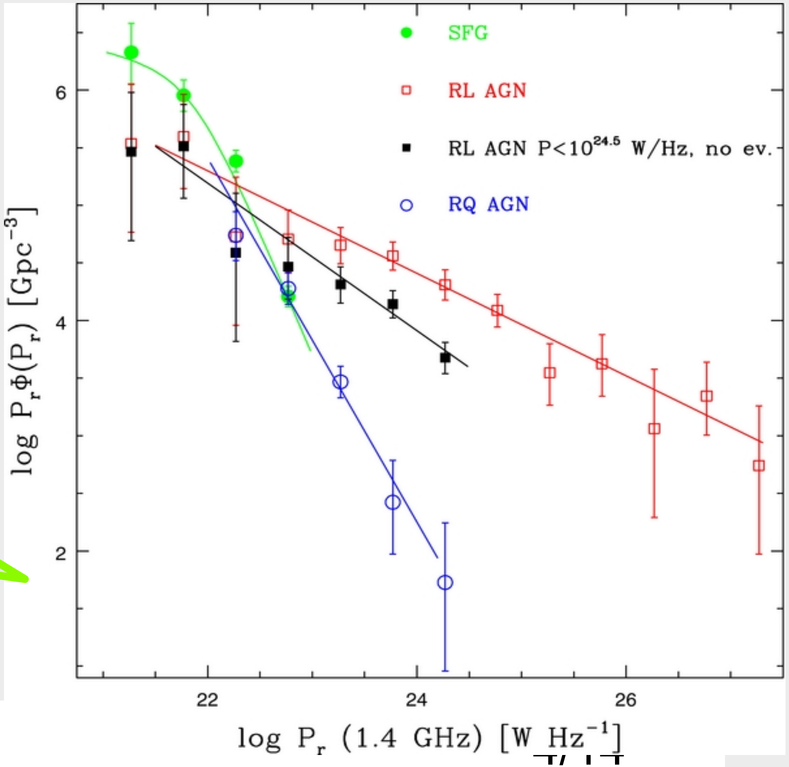




Why bother?

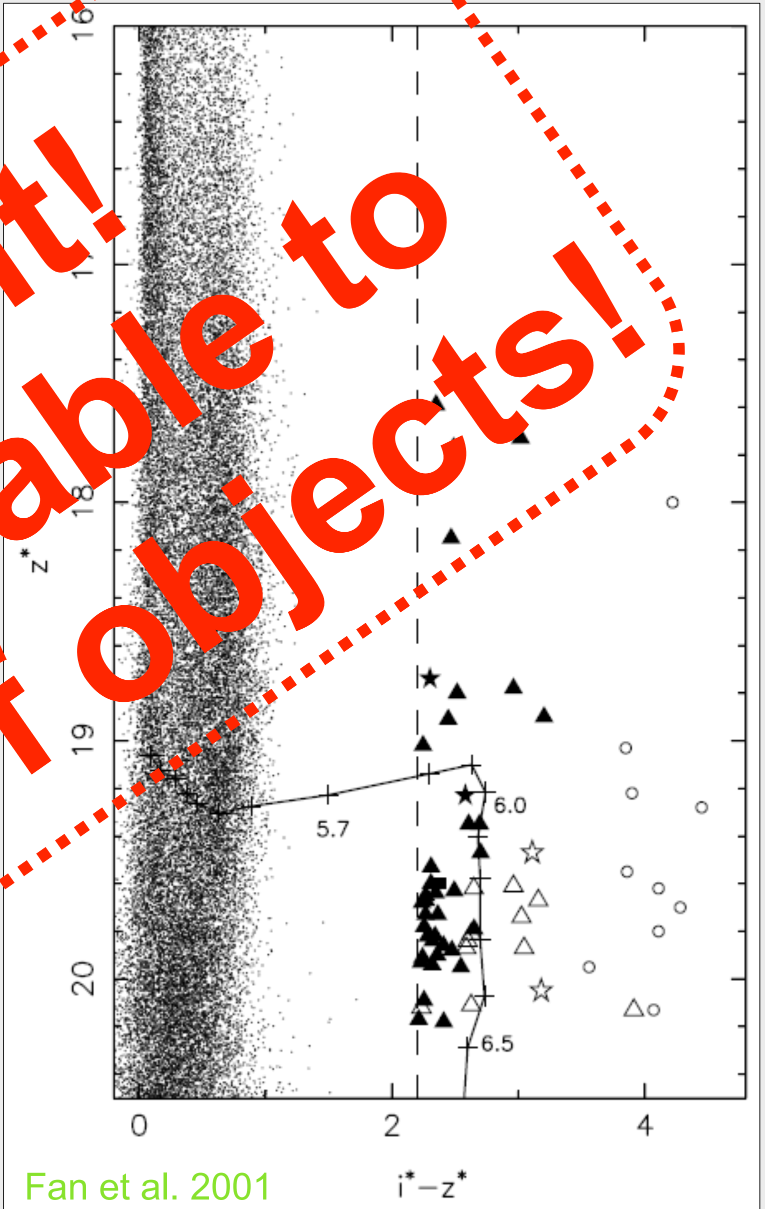
- Exact classifications and (at least) coarse redshifts are crucial for a large variety of science cases:
  - Co-evolution of AGN and their hosts
  - Most cosmology stuff
  - The cosmic star formation history
  - Luminosity functions / number counts
  - The radio/FIR correlation
  - …

Example: Padovani et. Al (2011)
→ LFs for RL & RQ AGN and SFGs
→ RQ AGN resemble SFGs



Common approaches

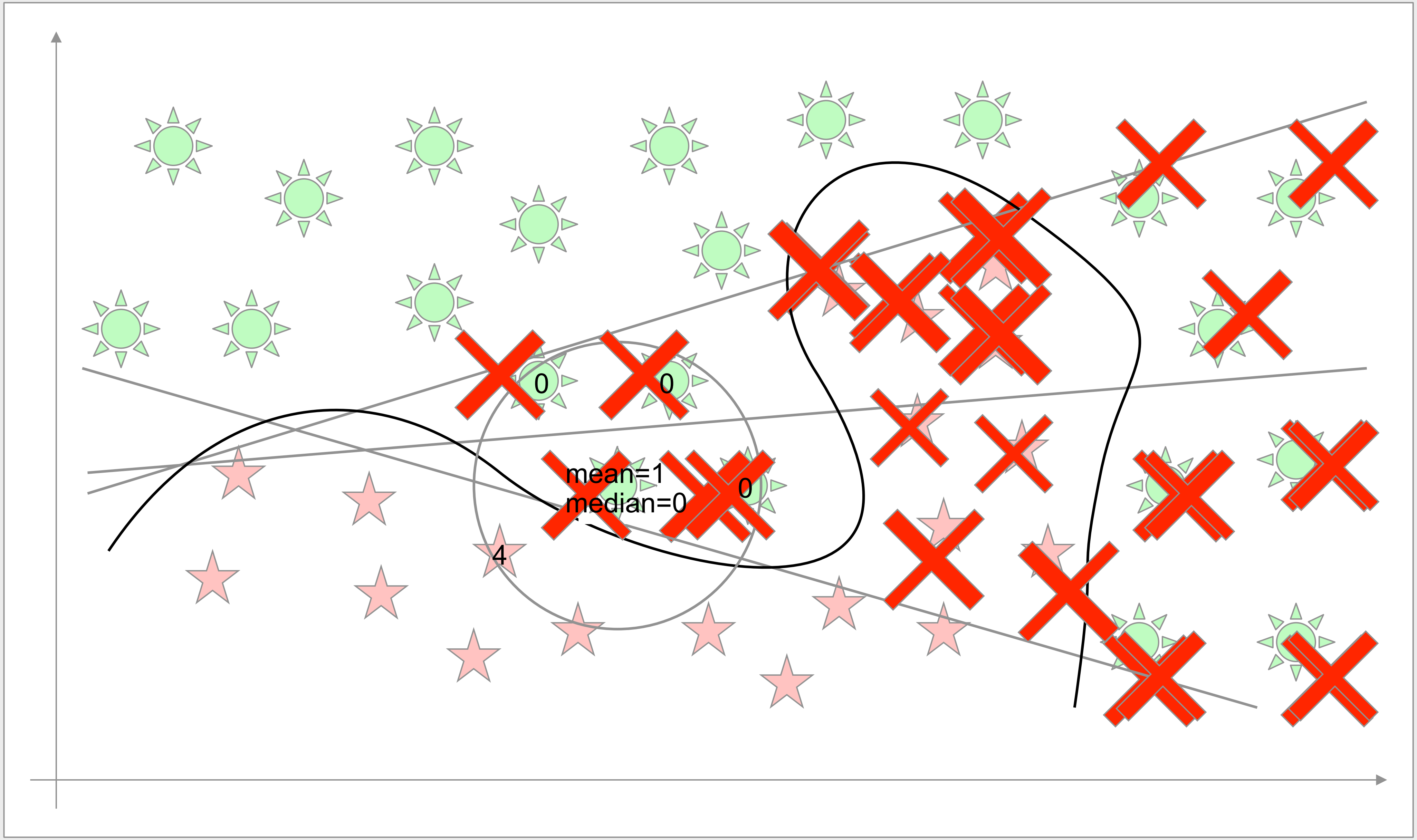
- Define plain color criteria
- Model SEDs
- Look for morphology, scaling relations, ...
- PROs:
  - Well-known → lots of expertise
  - Easy to understand for humans (2-dimensional selection criteria)
- CONs:
  - Global models → one number must fit everything
  - Hardly applicable to high-dimensional data sets
  - Require massive pre-processing
  - Most criteria physically motivated



Fan et al. 2001

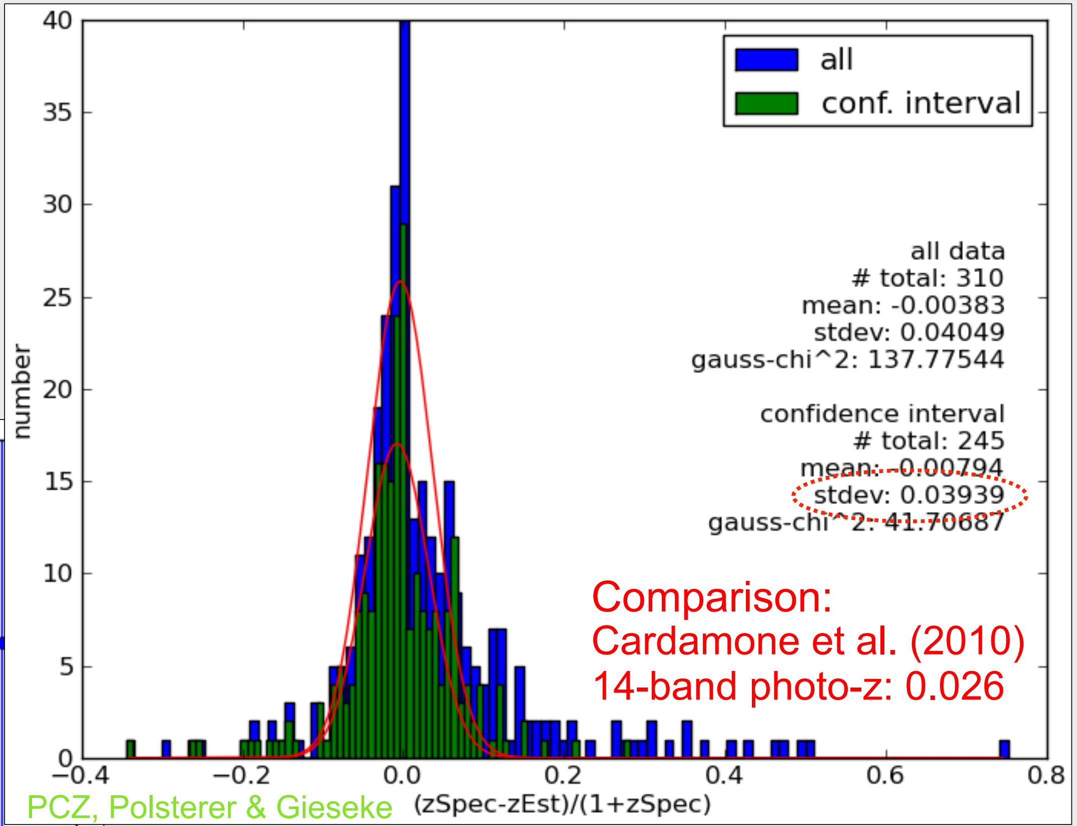
Inefficient!
Not applicable to
millions of objects!

Our approach: k nearest neighbors



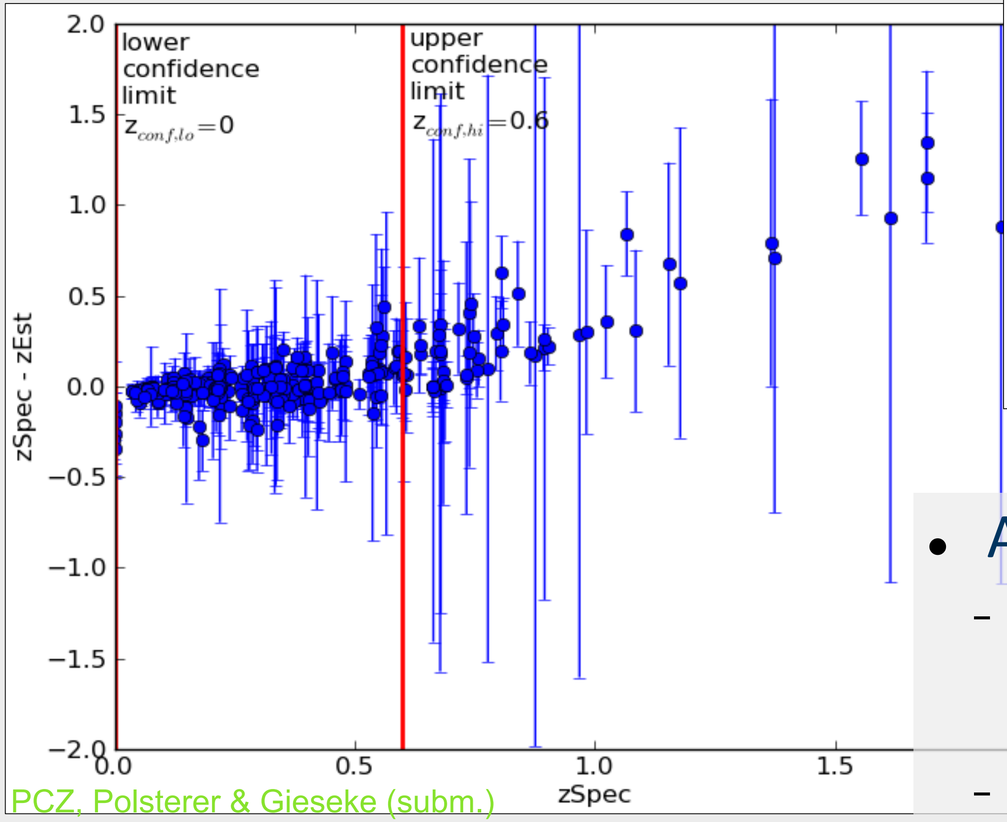
Example 1: kNN redshifts

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



Comparison:
Cardamone et al. (2010)
14-band photo-z: 0.026

PCZ, Polsterer & Gieseke

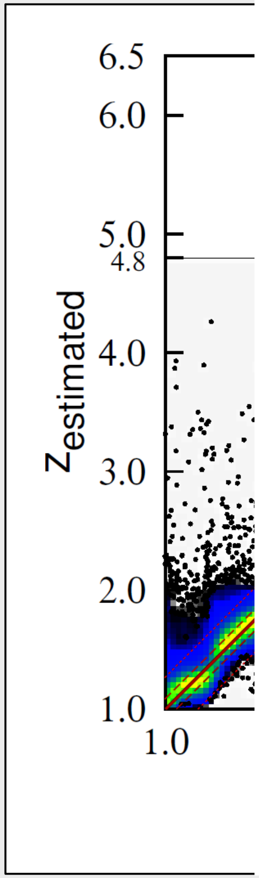


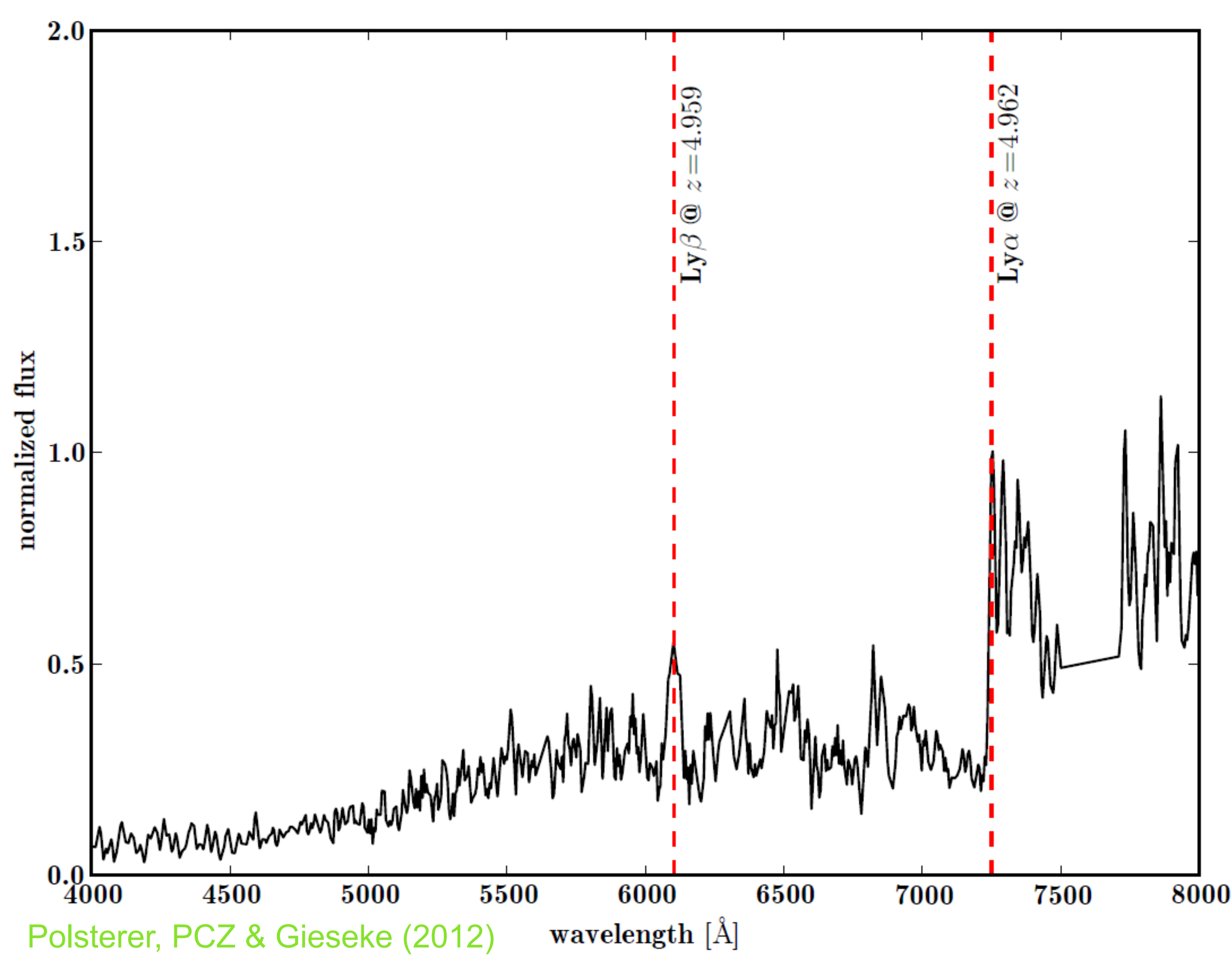
PCZ, Polsterer & Gieseke (subm.)

- Advantages of statistical redshifts
  - **No assumptions** must be made (no template SEDs, luminosity range, dust reddening, flux homogenization, ...)
  - Computation **much faster** than for class. photo-z ($t_{stat-z} \sim n*\log_2(n)$ | $t_{photo-z} \sim n^{\alpha}$, $\alpha>2$)

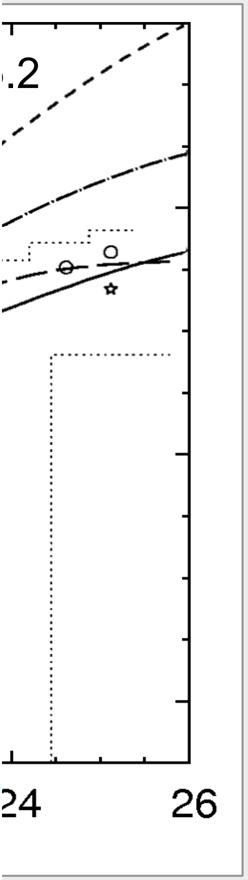
Redshift estimation for SDSS quasars

- kNN regression model + optimized training set
  - opti
  - 4-st fine class

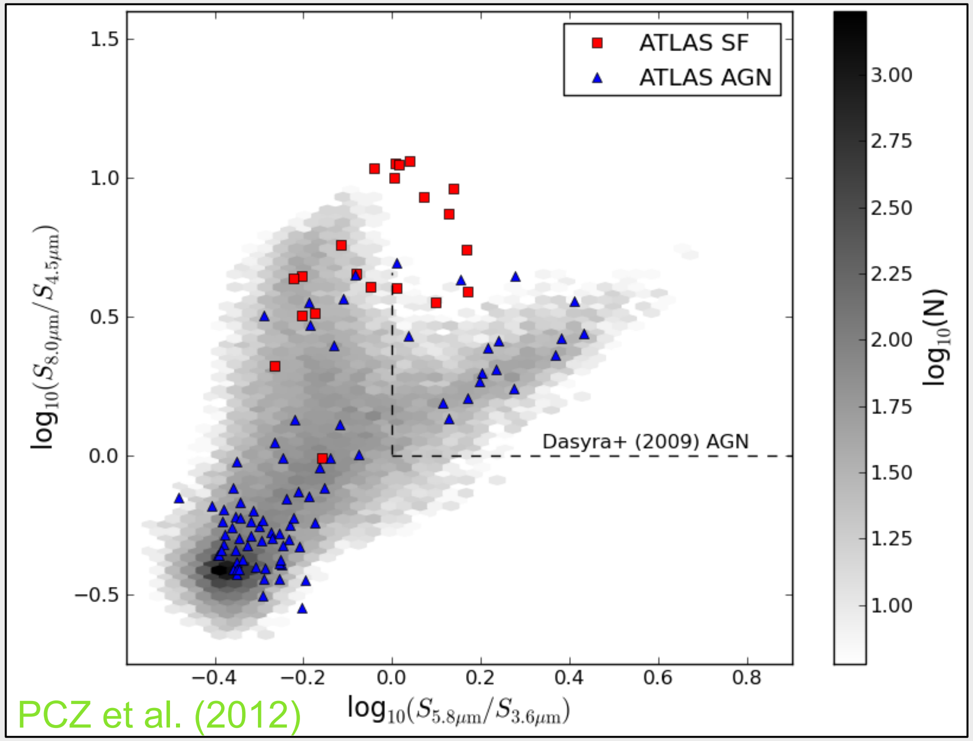




Polsterer, PCZ & Gieseke (2012)



Example 2: object classification

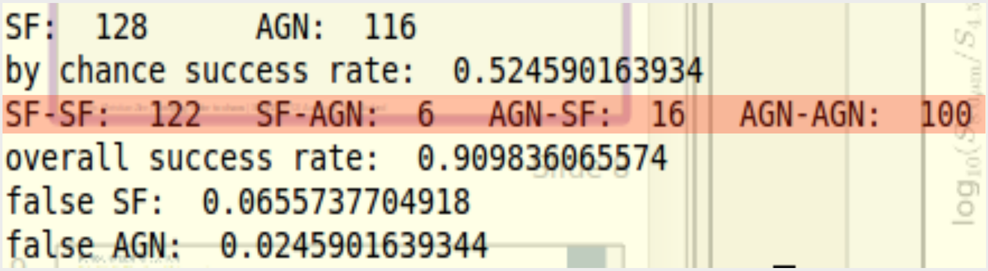


PCZ et al. (2012)

- SF / AGN separation
  - Classical tool: **BPT-diagram** (requires spectroscopy)
  - Alternative: **MIR color-color selection** (not very reliable)
  - **SED fitting** (work-intensive)

kNN-based classification of ATLAS test-sample yields combined **false classification rate of 9%**

Smolcic et al. (2008) achieve **contamination rates between 15% - 20%** using a highly sophisticated photometric method

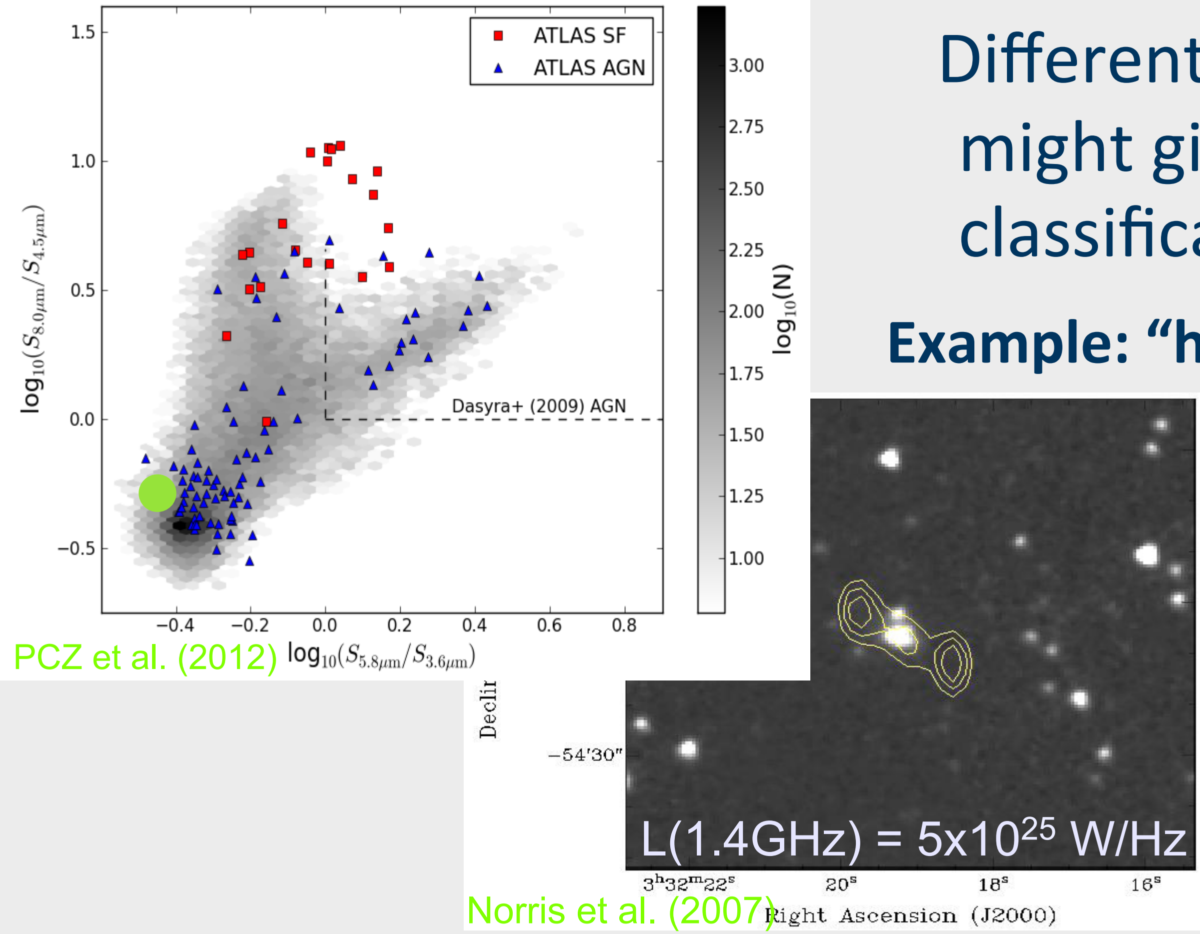


```
SF:  128      AGN:  116
by chance success rate:  0.524590163934
SF-SF:  122   SF-AGN:  6    AGN-SF:  16    AGN-AGN:  100
overall success rate:  0.909836065574
false SF:  0.0655737704918
false AGN:  0.0245901639344
```

Classification obstacles

Different classification methods might give you different classifications!

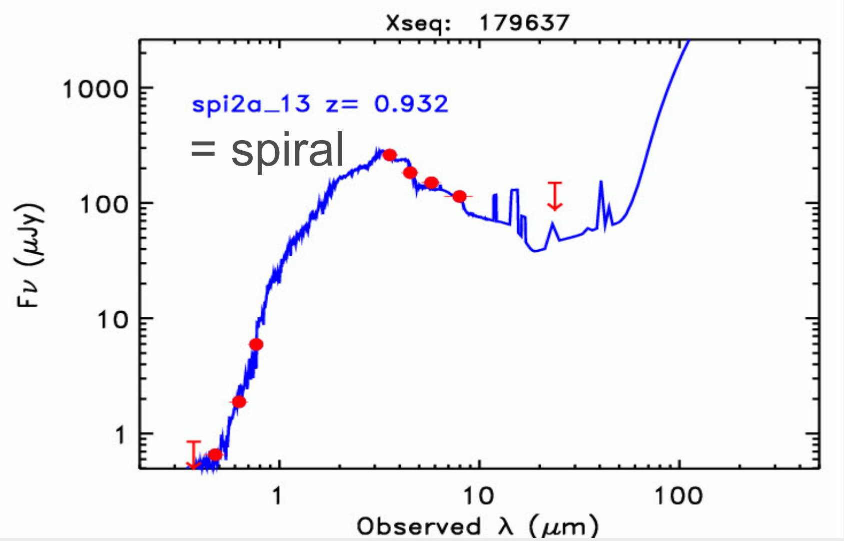
Example: “hidden” AGN



PCZ et al. (2012)

L(1.4GHz) = 5x10^{25} W/Hz

Norris et al. (2007)



spi2a_13 z= 0.932

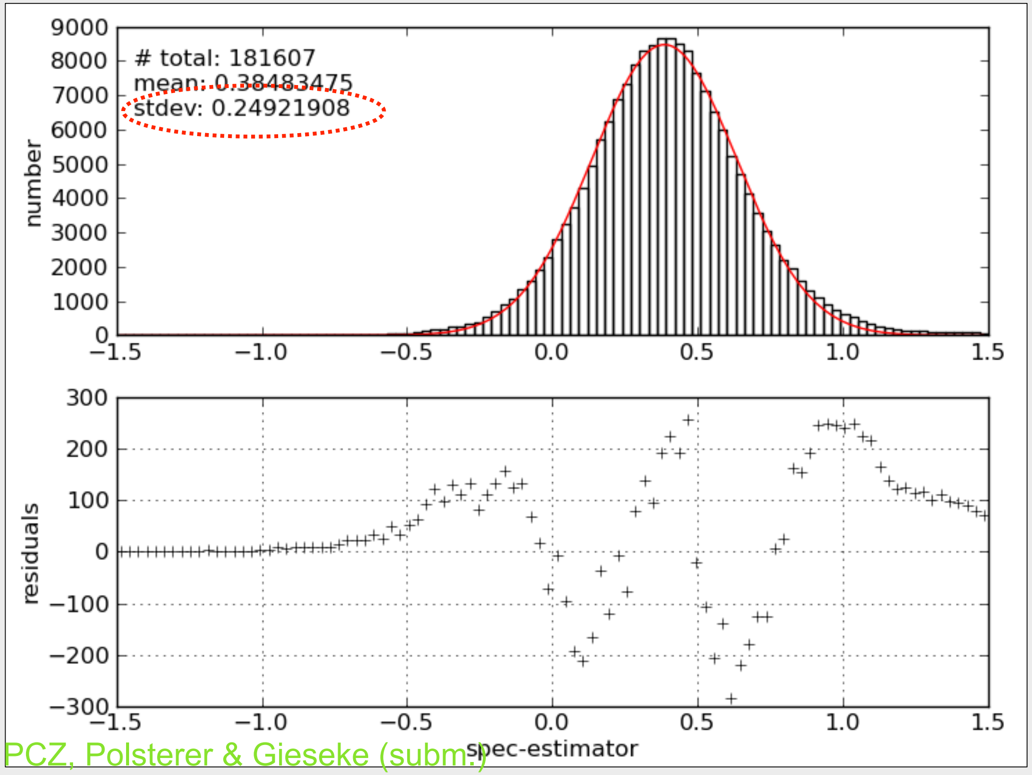
= spiral

- *Astrophysical obstacle:*
  At high redshift, AGN activity and star formation are closely linked (e.g. Mullaney et al. 2012, Rovilos et al. 2012, PCZ et al. in prep.)
- *Economical obstacle:*
  cross-matches for the entire electromagnetic spectrum needed

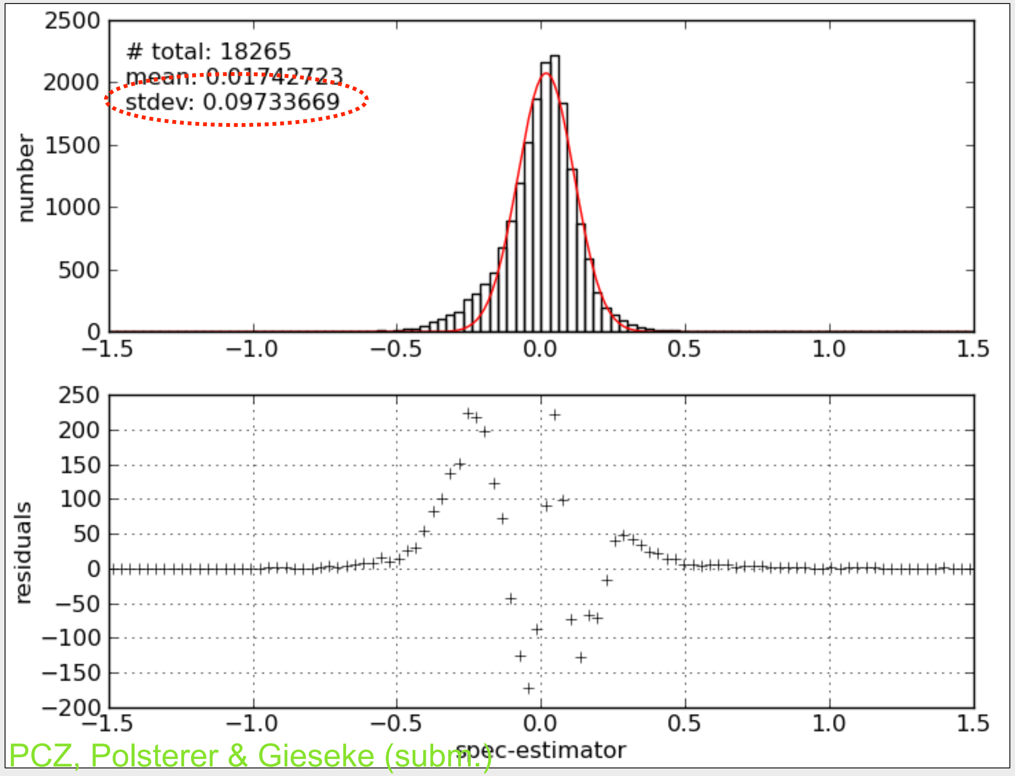
Example 3: metallicity

- 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 metallici-ties derived by Brinchman+ (2004)
  - kNN regression with respect to the 90 nearest neighbors
  - No other assumptions made



PCZ, Polsterer & Gieseke (subm.)



PCZ, Polsterer & Gieseke (subm.)

Example 4: stock market

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 surveys will produce.
- A **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**.
- Data-mining will already play an important role in currently upcoming projects, e.g. **ASKAP/EMU**.