Processing Large Data Sets: The Hunt for High-z Quasars

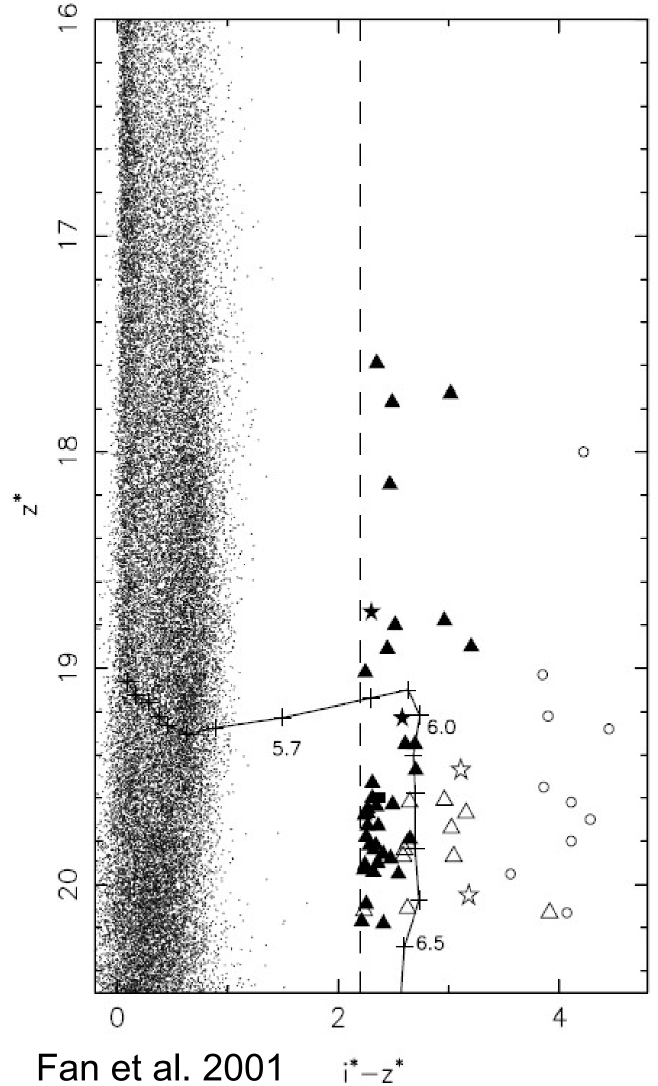
Kai Lars Polsterer, Peter-Christian Zinn and Fabian Gieseke

The Question

- can we efficiently find high-z quasars (z>4.8)
  - small computer / high prediction quality
- SDSS/DR6 catalogue was used
  - $300 \cdot 10^6$ objects observed in 5 filters (u,g,r,i,z)
  - $1 \cdot 10^6$ objects have spectra
  - $1 \cdot 10^5$ of these objects are known quasars
  - 150 of these quasars have z>4.8
  - covering 10,000 deg^2 (background image: 0.14 deg^2)

Common Approaches

- define plain colour criteria
  - PROs:
    - physically motivated
    - easy to reproduce in 2d diagrams
    - high completeness
  - CONs:
    - global model
    - does not work for high dimensions
    - many false positive candidates



Fan et al. 2001

Our Approach

- use k-Nearest Neighbours
  - local model
  - works fine in high dimensions
  - does not require physical assumptions
  - good reference samples available

$$\forall t_n \epsilon T, \hat{R}_{t(\overrightarrow{x})=t_n} = \frac{1}{k} \sum_{\overrightarrow{x}_i \epsilon N_k(\overrightarrow{x})} \begin{cases} 1, & t_i = t_n \\ 0, & \text{otherwise} \end{cases}$$

Finding the Nearest Neighbours

- neighbourhood search in Euclidean space
  - look-up implemented with kd-tree
- new distance to deal with measurement errors

$$d(\overrightarrow{u}, \overrightarrow{\Delta u}, \overrightarrow{v}, \overrightarrow{\Delta v}) = \sum_{i=1}^{N} \frac{(u_i - v_i)^2}{\Delta u_i^2 + \Delta v_i^2} + (|\Delta u_i| - |\Delta v_i|)^2$$

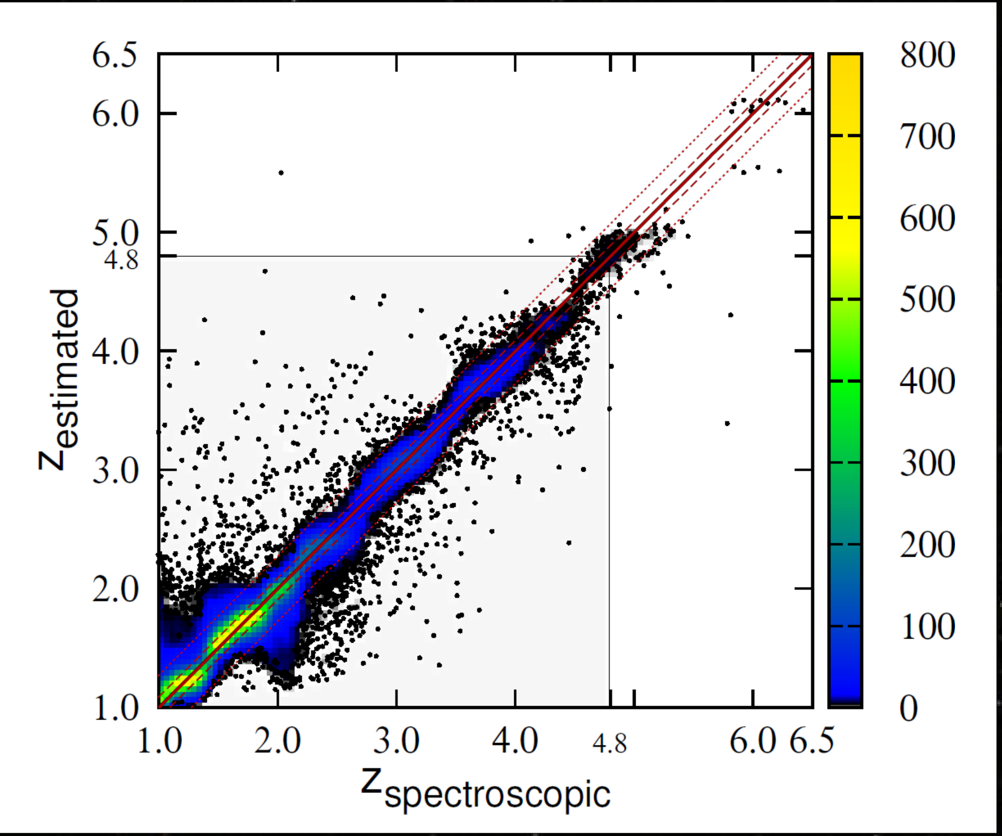
Classification

- 2 reference sets have been created
  - first reference set
    - all 1,258 z>4 + 1,000 medium redshift quasars
    - 1,000 galaxies + 1,000 stars + 1,500 cool stars
  - second reference set
    - all 1,258 z>4 quasars
    - 10,900 cool stars
- neighbours are stored
  - ratios can be calculated later

Redshift Estimation

- kNN regression model + selected reference set
  - 77,000 references reduced to 1,100 objects
  - optimised for z > 4.8
  - 4 colours used

$$\hat{Y}(\vec{x}) = \frac{1}{k} \sum_{\vec{x}_i \epsilon N_k(\vec{x})} y_i$$



Candidate Selection

- 4 rejection filters combined
  - coarse / redshift / cool stars / new distance
- optimised for speed
  - 1 hour / 1000 objects with first implementation
    - 37.7 years on one core
  - 2-8 seconds / 1000 objects with optimisation
    - efficient data structures
    - optimised reference sets
    - parallel execution
    - 14 hours on 8 cores

Results

- ratios optimised with all SDSS objects with spectra
  - 50% of all known high-z quasars are recovered
  - 40% are false positives
  - only 0.1% of the cool stars pass the rejection stage
- 122,000 candidates are returned

The Answer

- 3 candidates observed
  - with SCORPIO @ 6m BTA