

# Robust Machine Learning Applied to Astronomical Datasets: Photometric Redshifts

Nick Ball

Department of Astronomy and National Center  
for Supercomputing Applications

University of Illinois at Urbana-Champaign

DES Collaboration Meeting, Chicago, Dec 12th 2006

# Collaborators

- **Laboratory for Cosmological Data Mining (LCDM)** at **NCSA** and **UIUC Astronomy**: Robert Brunner, Adam Myers, Natalie Strand, Stacey Alberts
- **Automated Learning Group, NCSA**: David Tcheng, Xavier Llorà
- LCDM is a top-20 user of NCSA supercomputing resources



alg



# Photozs: Quasars, Galaxies

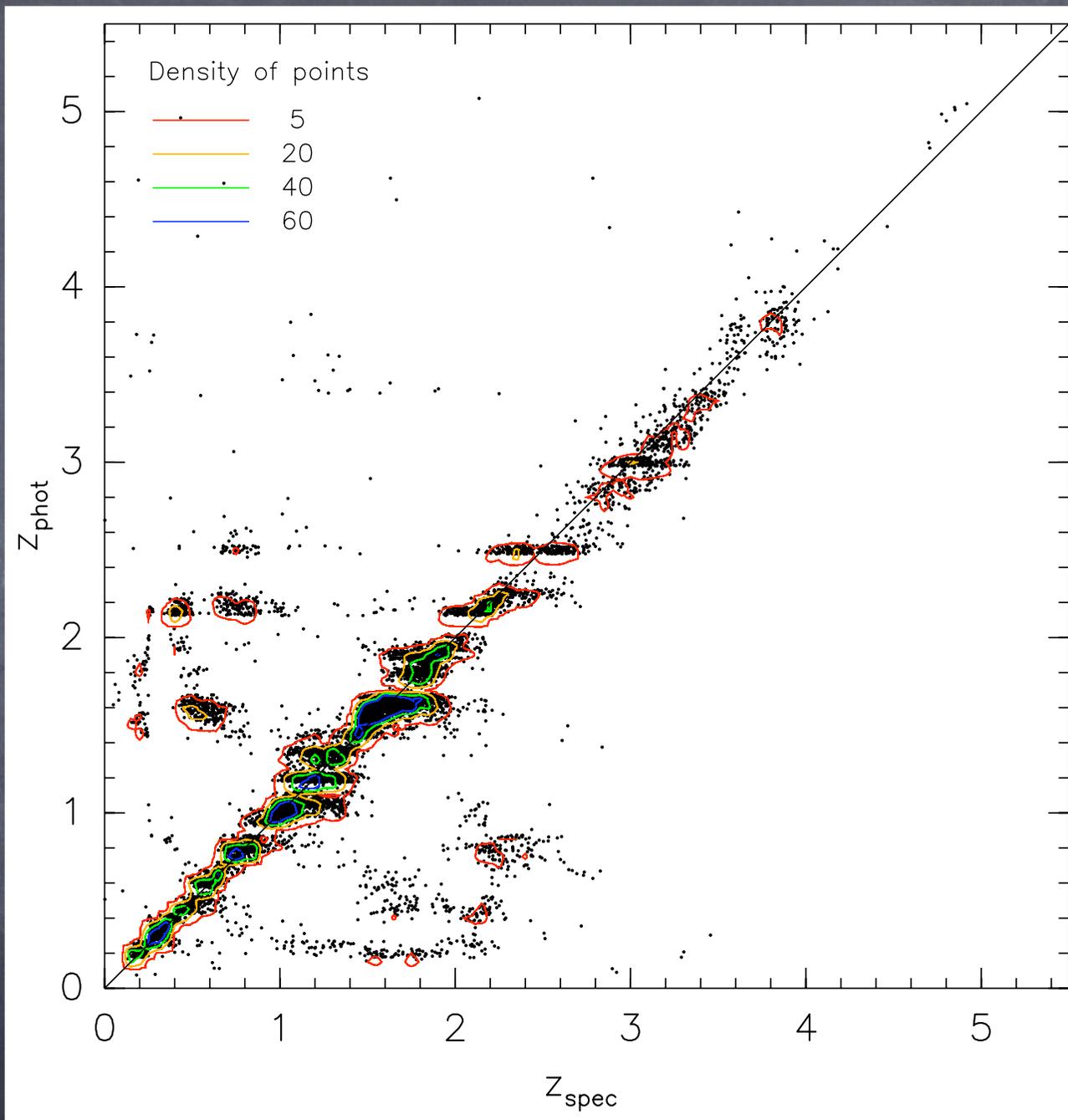
- We apply instance-based learning to obtain photometric redshifts for objects in the **SDSS DR5** and **GALEX GR2**
- We use the Java environment **Data to Knowledge** and the NCSA Xeon Linux supercomputing cluster **Tungsten**
- Here we present results for quasars, then preliminary results for galaxies

# Instance-Based Learning

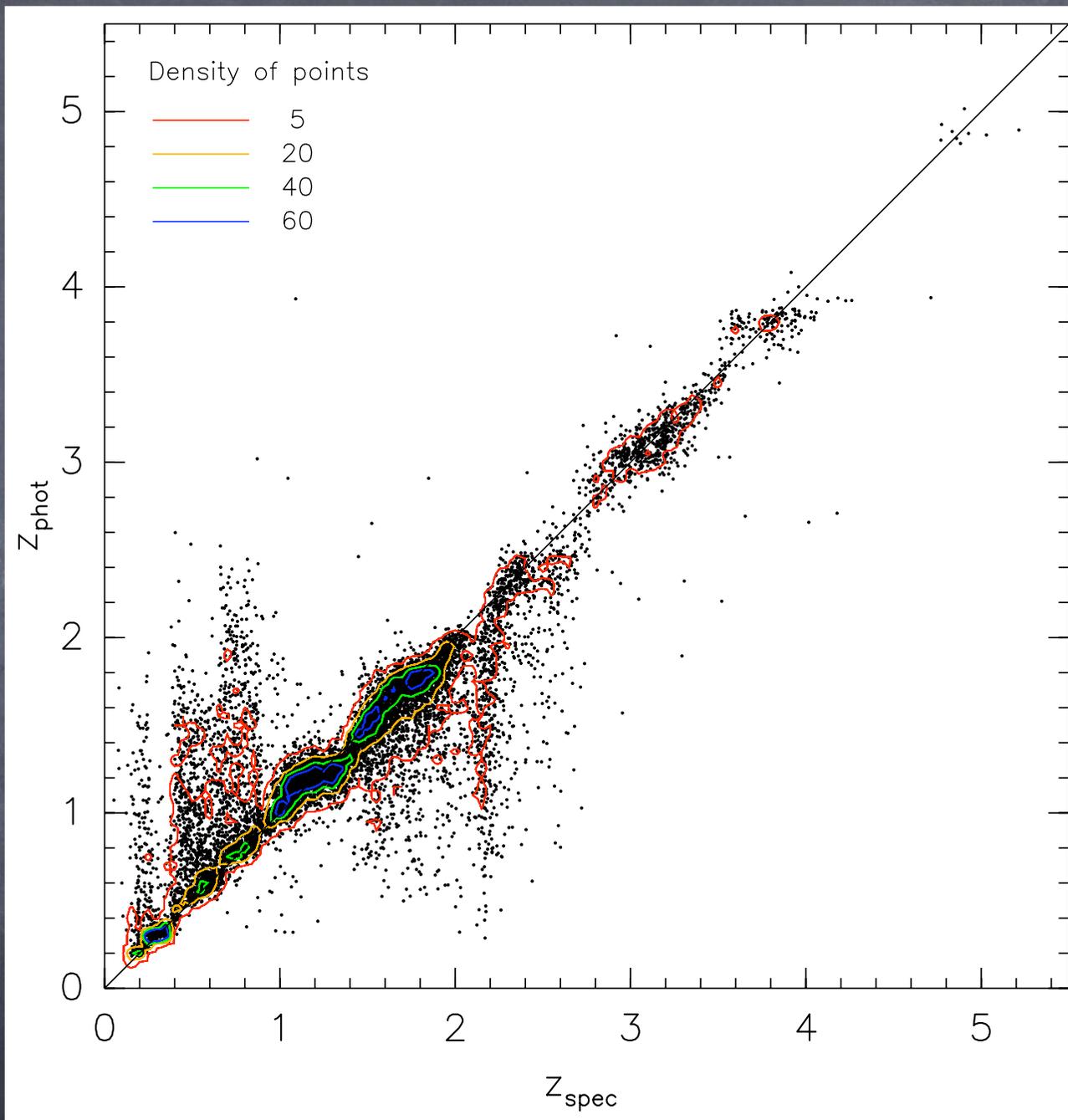
- Memorize the positions in parameter space of each training object
- For new objects, calculate the weighted average redshift of the  $k$  nearest neighbors
- Most of the work is done in the latter stage
- Computationally intensive

# Quasar Photozs

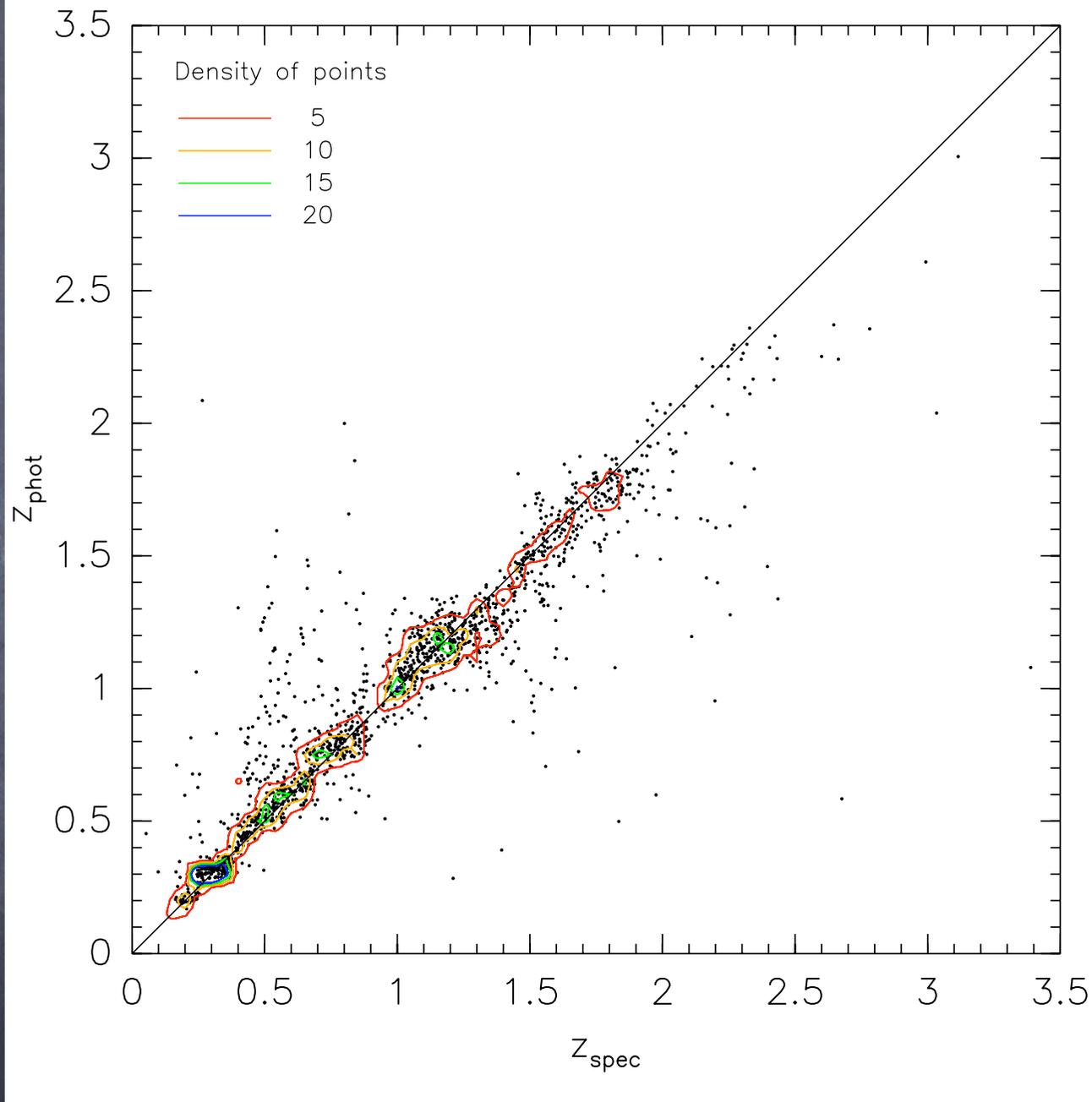
- We assign photozs to 55,746 SDSS DR5 quasars and 7,642 SDSS DR5+GALEX GR2 quasars ( $i < 19.1$ )
- We use a CZR and compare it to instance-based learning
- We train on 80% and blind test on 20%
- This gives blind testing samples of 11,149 for SDSS and 1,528 for SDSS+GALEX



SDSS CZR blind test: 11,149 of 55,746 quasars



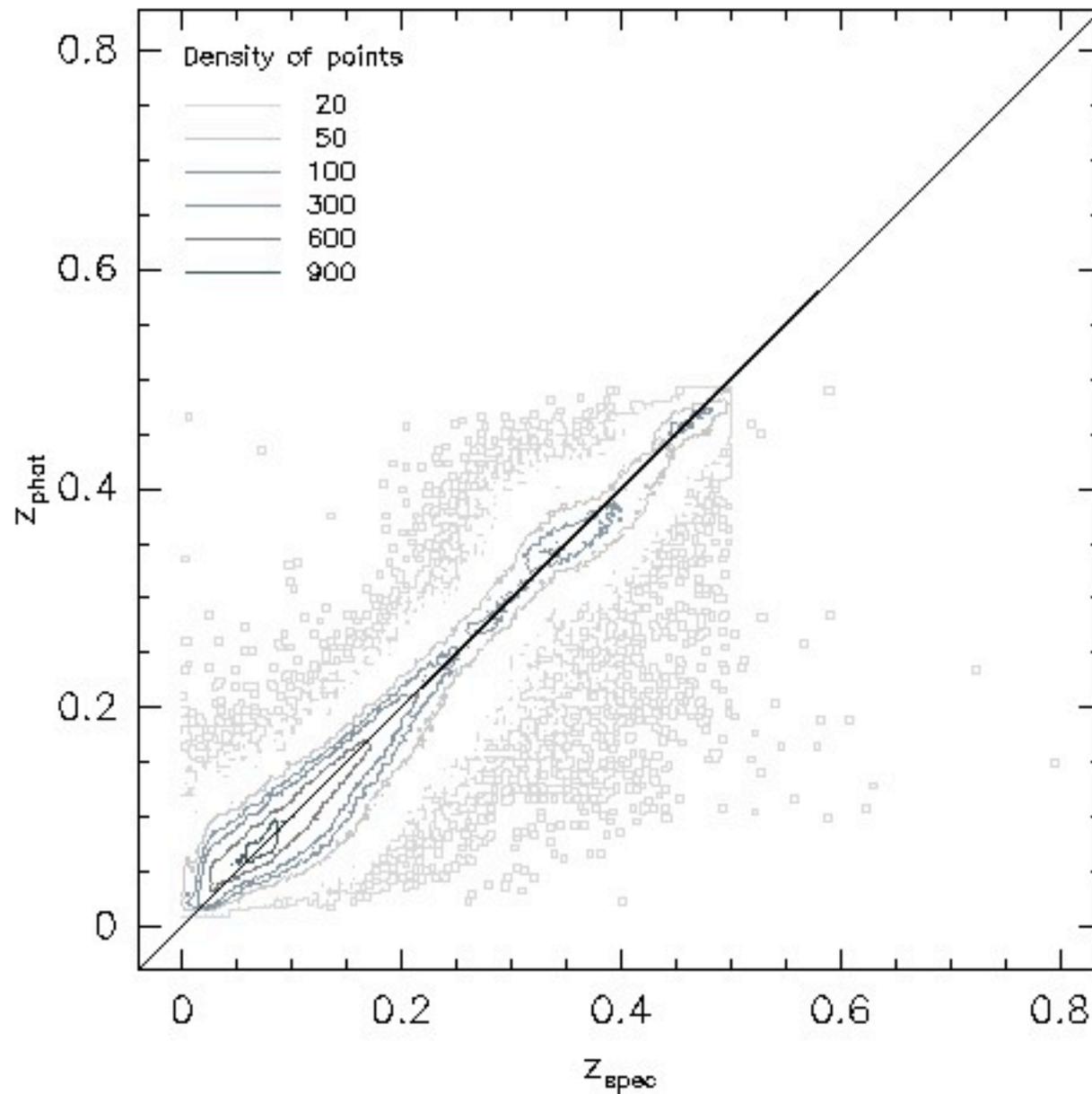
SDSS k-NN instance-based blind test:  
11,149 of 55,746 quasars



SDSS+GALEX k-NN instance-based blind test:  
1,528 of 7,642 quasars

# Galaxy Photozs

- We have assigned preliminary galaxy photozs to **SDSS DR5 Main** galaxies ( $r < 17.77$ ) using a decision tree
- The RMS dispersion is **0.02**
- This is similar to existing photozs for these galaxies



SDSS DR5 Main galaxies

# Next Steps

- Full PDFs incorporated into the machine learning and output photozs
- Assign photozs with PDFs to 200 million objects in SDSS photoPrimary, as done for classification into star-galaxy-neither by Ball et al. 2006a (ApJ 650 497)
- Use of (funded by NASA AISR) High Performance Reconfigurable Computing (HPRC) in collaboration with NCSA Innovative Systems Laboratory
- Further multiwavelength training data

# Conclusions

- We have assigned photozs to quasars and in the SDSS DR5 and GALEX GR2
- We have assigned preliminary photozs to SDSS DR5 Main galaxies
- We find that instance-based learning reduces the incidence of catastrophic failures in quasar photozs compared to CZR

• <http://nball.astro.uiuc.edu>

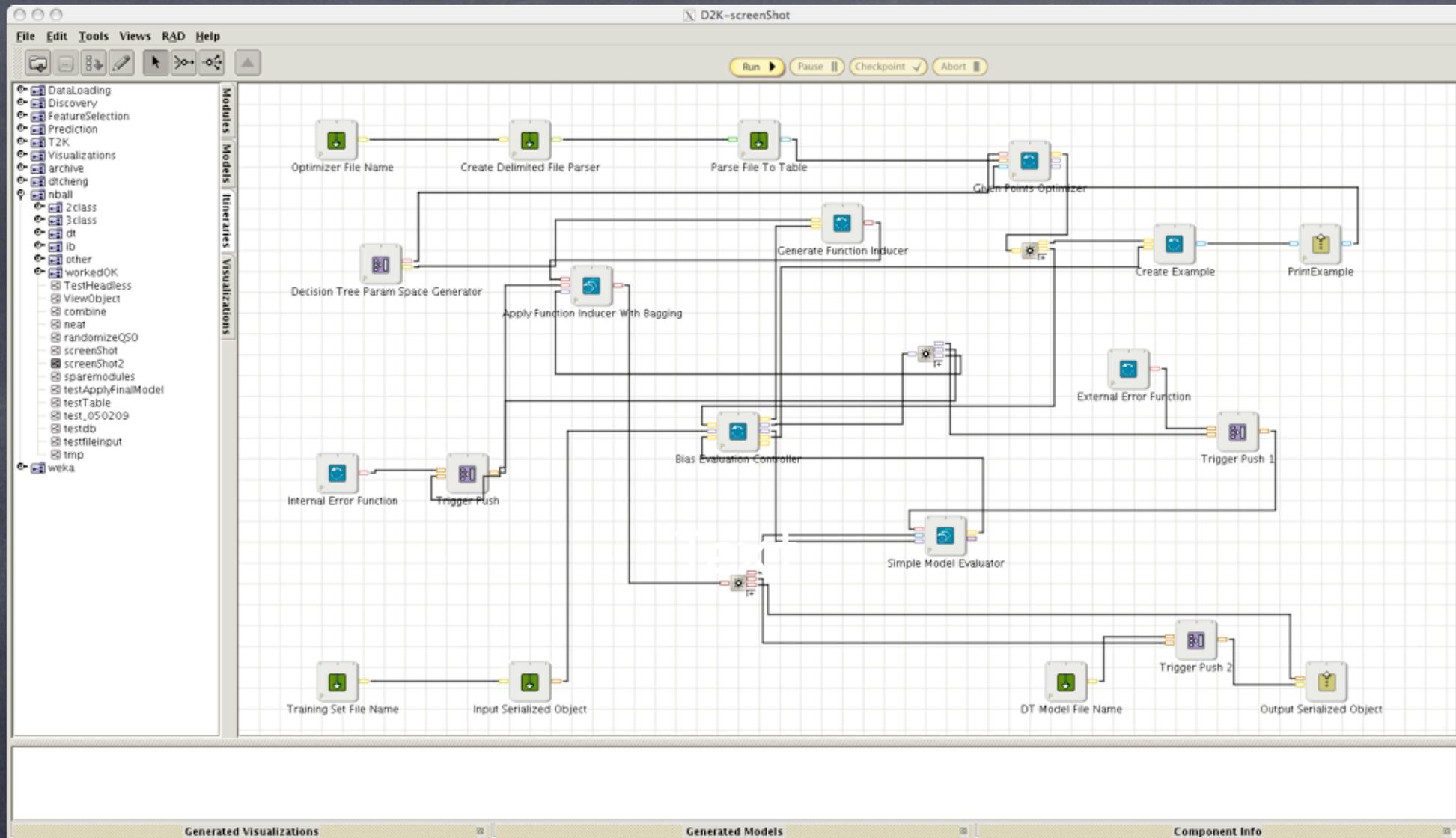
• Ball et al. 2006b, in preparation

• DES uploaded talks (extra slides)

Extra slides...

# D2K

- We use the Java environment **Data to Knowledge**, developed at NCSA
- Modified to run on multiple Tungsten nodes and multi-GB-sized datasets
- D2K itineraries automate the data-mining process
- Many different algorithms are available

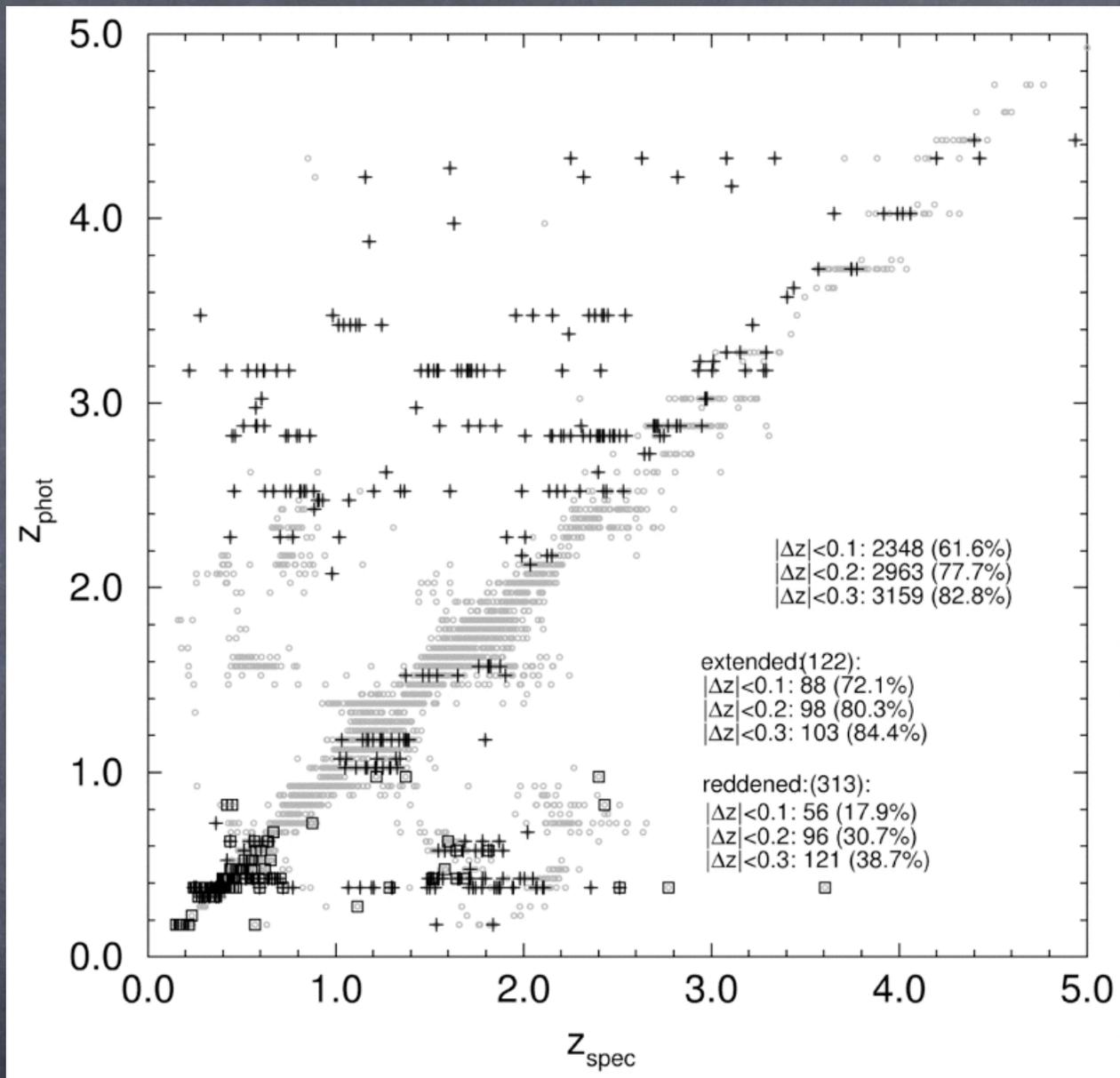


D2K screenshot

# NCSA Supercomputing

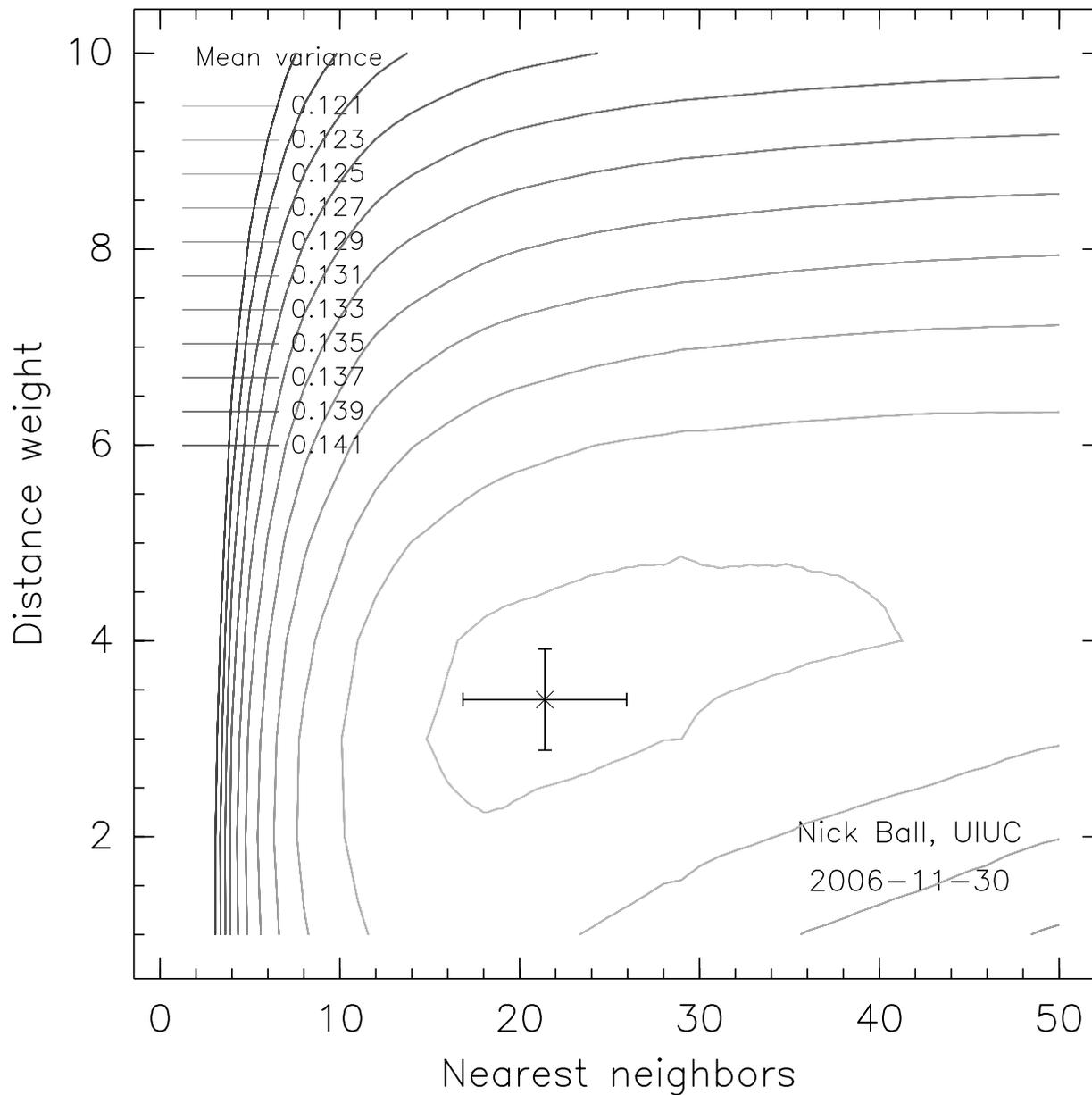
- Xeon Linux Cluster  
Tungsten
- 2,560 Intel IA-32  
Xeon 3.2 GHz  
processors, 3 GB  
memory/node
- Peak performance:  
16.38 TF (9.819 TF  
sustained)





CZR, 3,814 SDSS EDR quasars, Weinstein et al.  
2004 (ApJS 155 243)

SDSS DR5



Instance-based: effect of number of nearest neighbors and distance weighting

Dataset	Method	Variance	Variance / (1+z)	Mean $ \Delta z  / (1+z)$	% within $ \Delta z  < 0.1$	% within $ \Delta z  < 0.2$	% within $ \Delta z  < 0.3$
SDSS	IB	$0.120 \pm 0.002$	$0.034 \pm x.xxx$	$0.097 \pm x.xxx$	$55.1 \pm x.x$	$73.6 \pm x.x$	$80.5 \pm x.x$
SDSS +GALEX	IB	$0.058 \pm 0.004$	$0.014 \pm 0.002$	$0.060 \pm 0.003$	$70.8 \pm 1.2$	$85.8 \pm 1.0$	$90.8 \pm 0.7$
GALEX- SDSS-only	IB	$0.093 \pm 0.010$	$0.022 \pm 0.001$	$0.081 \pm 0.003$	$62.0 \pm 1.4$	$78.9 \pm 1.0$	$85.2 \pm 1.2$
SDSS	CZR	$0.265 \pm 0.006$	$0.079 \pm 0.003$	$0.115 \pm 0.002$	$63.9 \pm 0.3$	$80.2 \pm 0.4$	$85.7 \pm 0.3$
SDSS +GALEX	CZR	$0.136 \pm 0.015$	$0.031 \pm 0.006$	$0.071 \pm 0.005$	$74.9 \pm 1.4$	$86.9 \pm 0.6$	$91.0 \pm 0.8$
GALEX- SDSS-only	CZR	$0.158 \pm 0.013$	$0.041 \pm 0.004$	$0.081 \pm 0.004$	$74.1 \pm 0.8$	$86.2 \pm 0.7$	$89.7 \pm 0.6$