Deep Learning Astronomical Survey Data

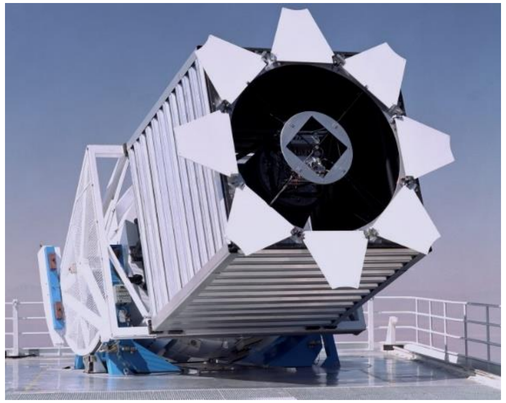
Shawfeng Dong

shaw@ucsc.edu

University of California, Santa Cruz

2017 DOMA Workshop, New York, NY

Challenges of Big Data in Astronomy



Sloan Digital Sky Survey (SDSS)

- 200 GB per night
- 40 TB raw data ➠ 120 TB processed
- 35 TB catalogs

Mikulski Archive for Space Telescopes (MAST)

- 185 TB of images
- 25 TB/year ingest rate
- 100 TB/year retrieval rate

Large Synoptic Survey Telescope (LSST)

- 15 TB per night for 10 years
- 100 PB image archive
- 20 PB final database catalog

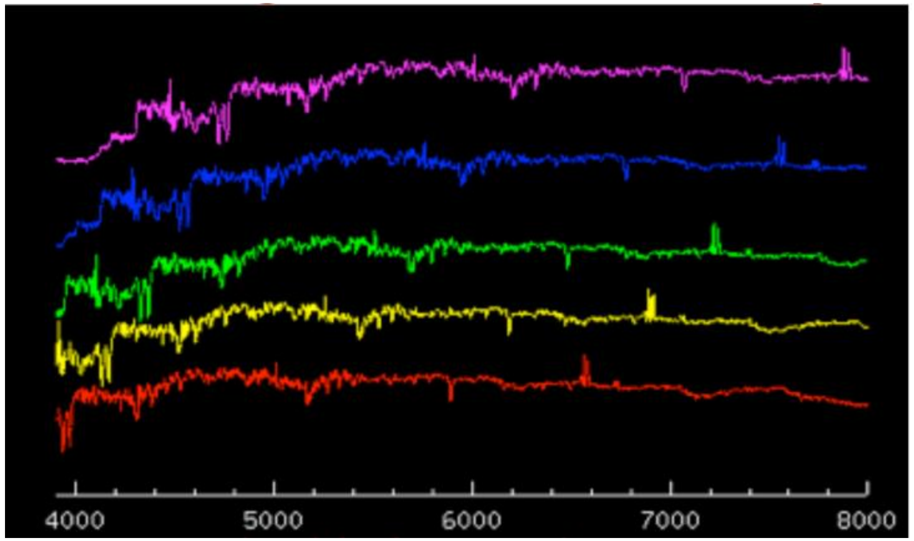
Square Kilometer Array (SKA)

- 1 EB per day (> internet traffic today)
- 100 PF processing power
- ~1 EB processed data per year

Sloan Digital Sky Survey
"The Cosmic Genome Project"

Imaging survey in 5 wavelength bands

Spectroscopic redshift survey



Massive Data

- 200 GB per night
- 40 TB raw data ➠ 120 TB processed
- 35 TB catalogs

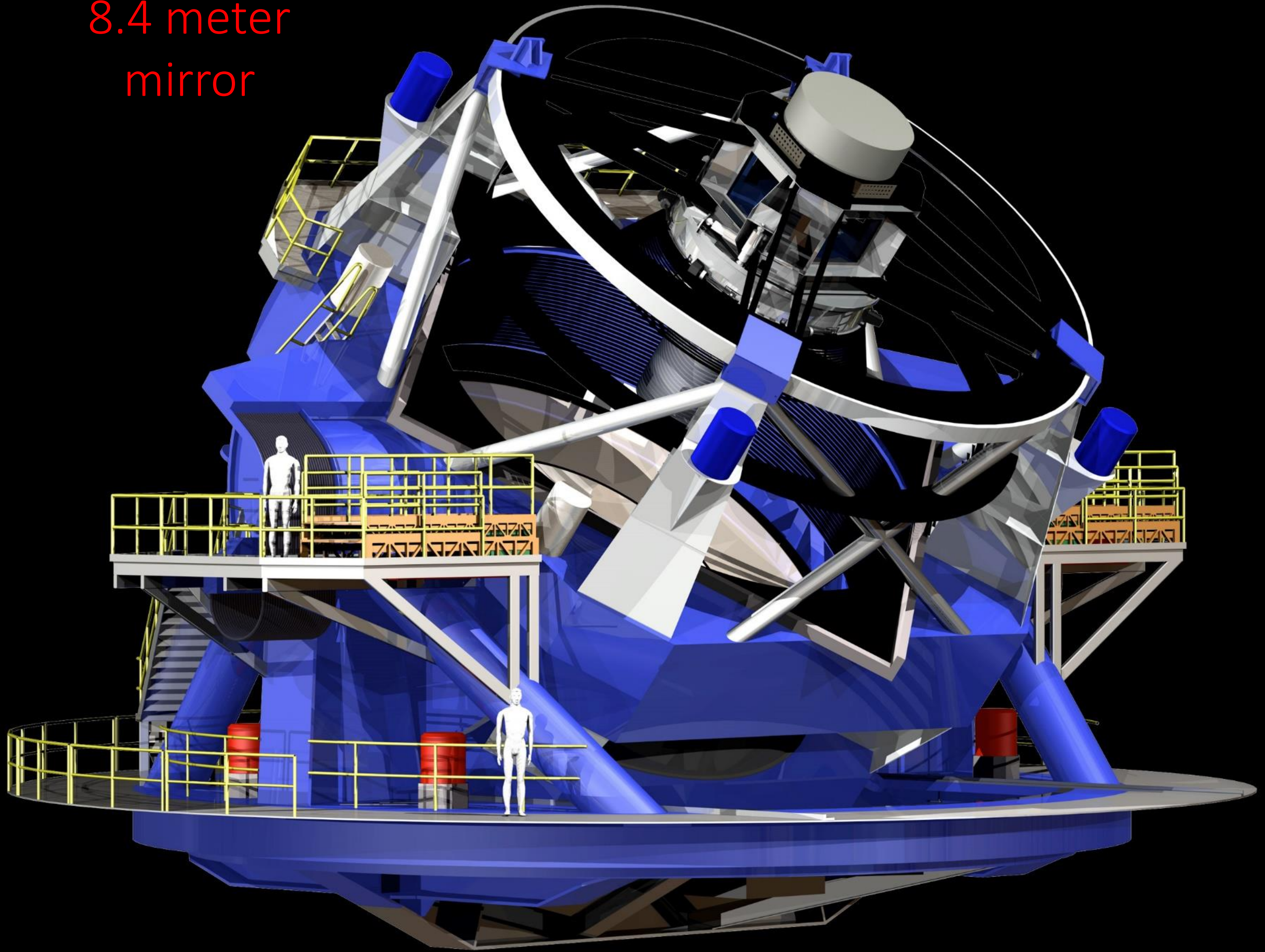
Data is publicly accessed

- 840 million web hits in 9 years, now >1 billion
- 4,000,000 distinct users vs. 15,000 astronomers
- Basis for >20,000 scientific papers
- More citations than any telescope including Hubble

SDSS Telescope (2.5 meter mirror)

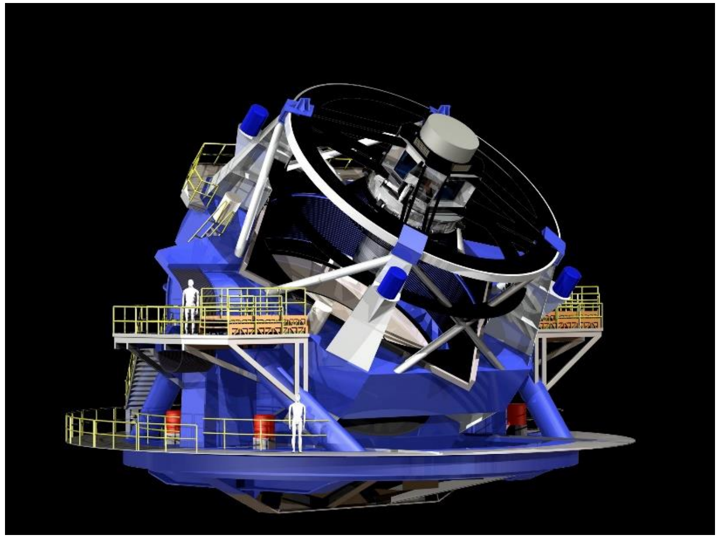
SDSS
Telescope
2.5 meter
mirror

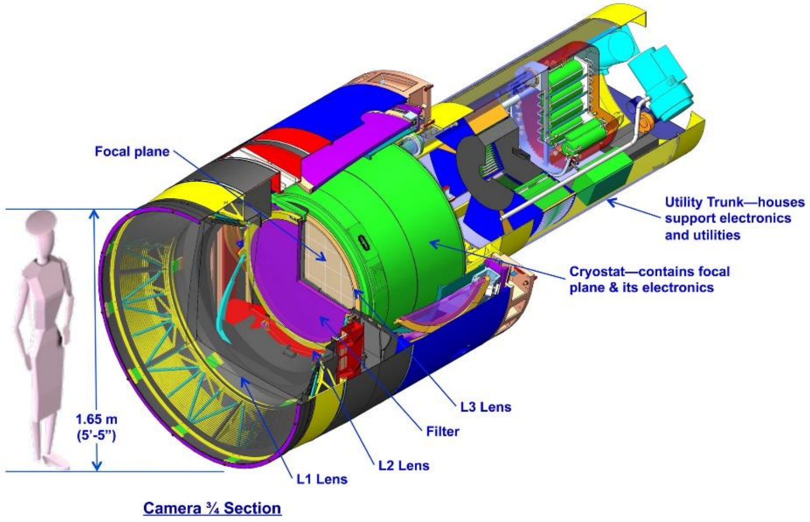
LSST
Telescope
8.4 meter
mirror



Large Synoptic Survey Telescope

- Wide field and deep
  - 27,000 square degrees (wide)
  - 100 – 200 square degrees (deep)
  - 10 years
- Broad range of science
  - Dark energy & dark matter
  - Galactic structure
  - Census of the Solar system
  - Transient universe
- 3.2-gigapixel camera
  - 9.6 square degrees FOV
  - 6 filters (UGRIZY)



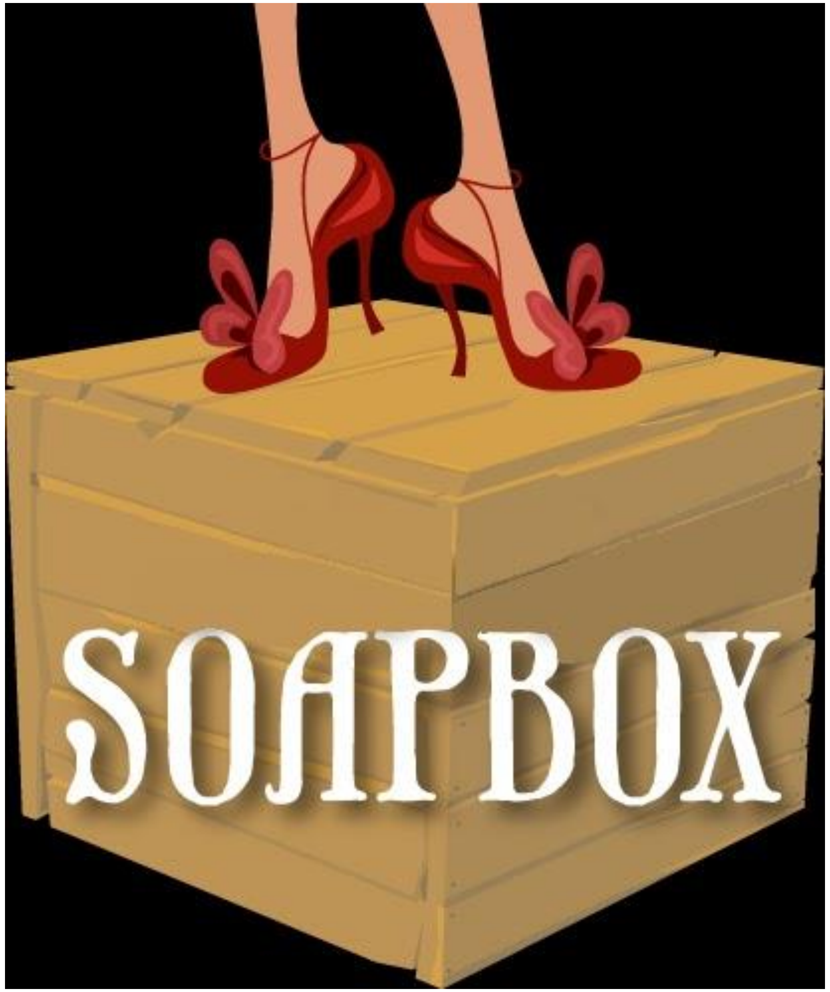


Processing the data flow from LSST

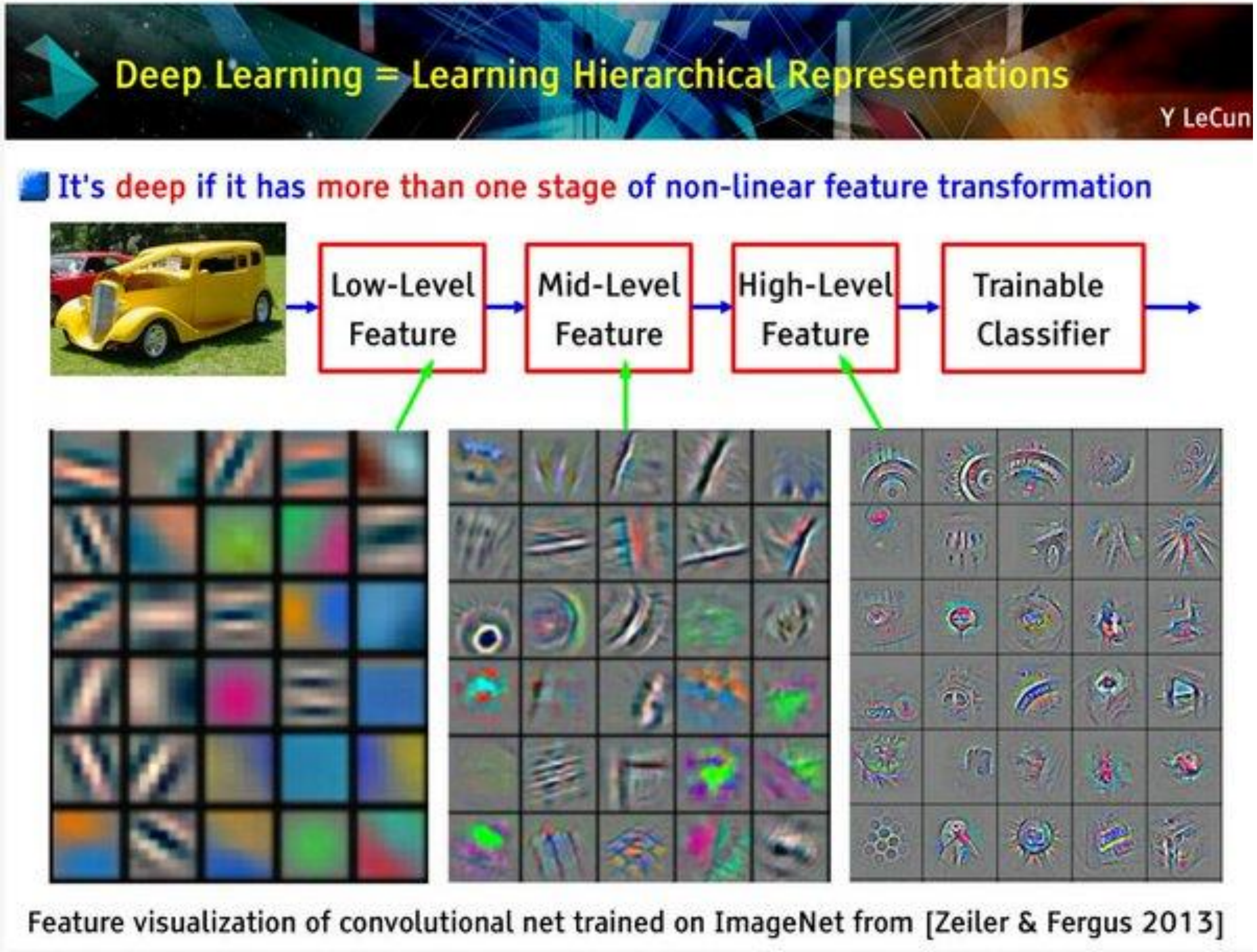
- A 15-second exposure every 20 seconds
- The data volume associated with this cadence is unprecedented!
  - one 6-gigabyte image every 20 seconds
  - 15 terabytes of raw scientific image data per night
  - 100-petabyte final image data archive
  - 20-petabyte final database catalog
  - 2 million real time events per night every night for 10 years
  - 1000 new supernovas discovered every night!
- Managing and effectively data mining the enormous output is expected to be the most technically difficult part of the project

Machine Learning in Astronomy

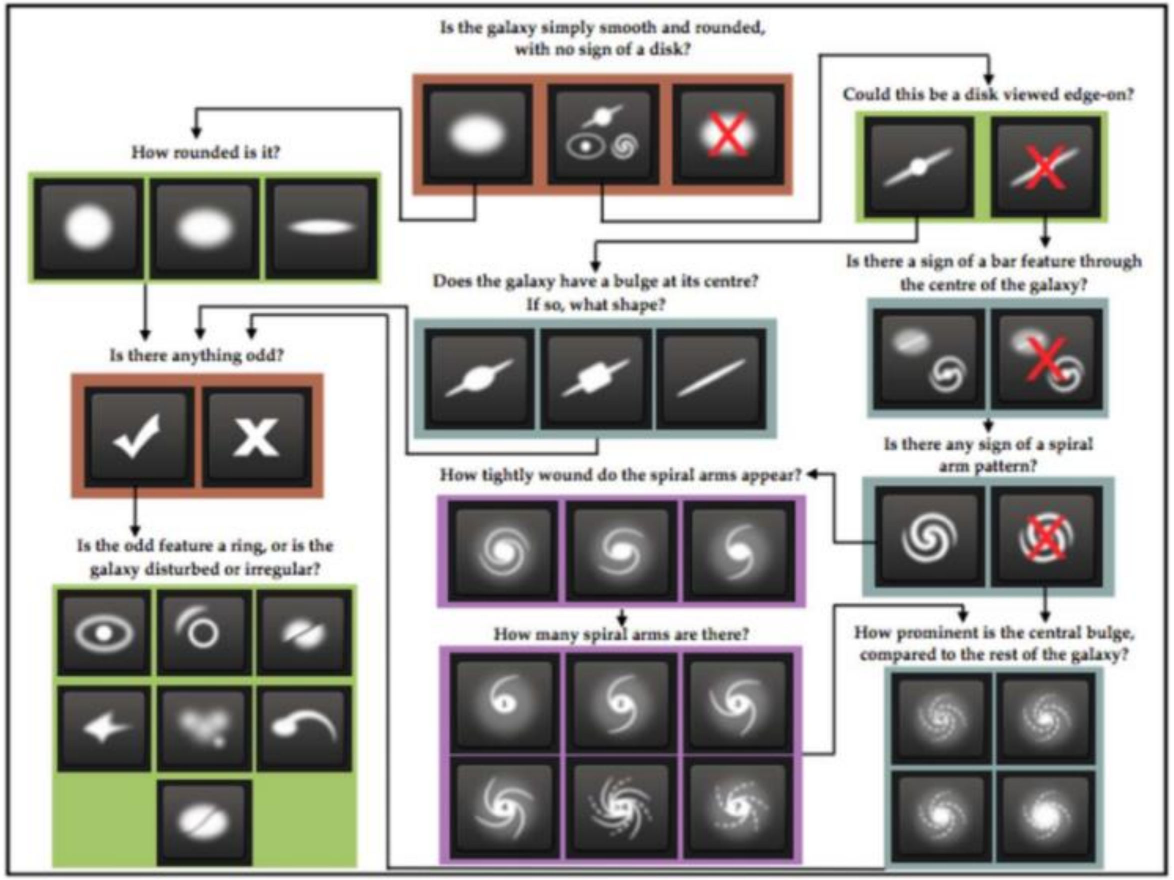
- Astronomy is rife with tasks demanding human labor
  - Source identification
  - Continuum fitting
  - Line identification
  - Etc.
- Machine Learning
  - Can perform many of these tasks
  - Auto-magically, repeatedly, better!



Deep Learning

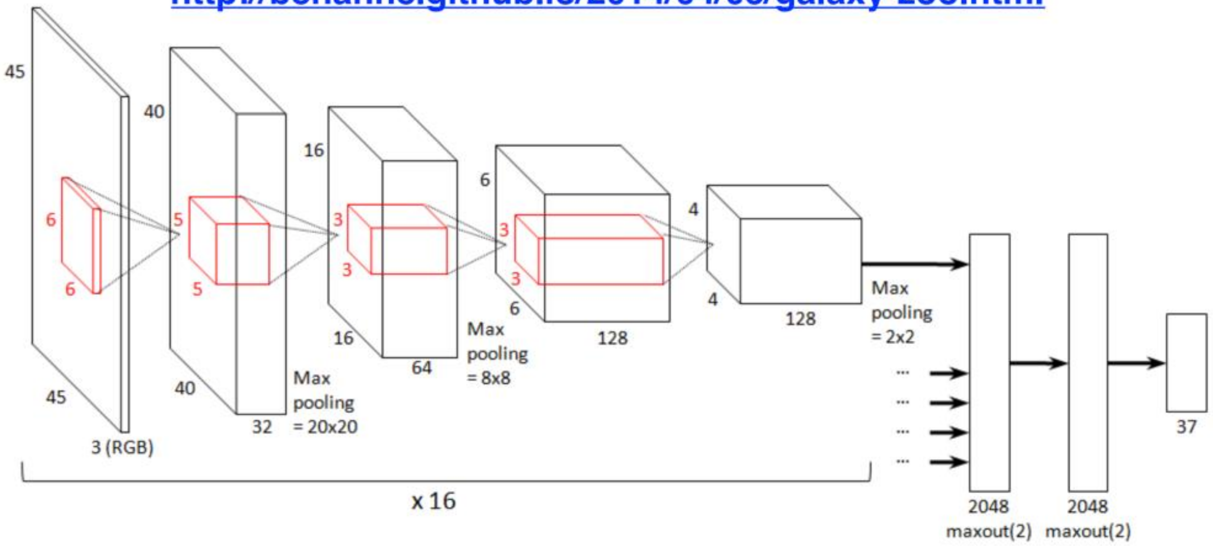


Sander Dieleman used deep learning to predict **Galaxy Zoo** nearby galaxy image classifications with 99% accuracy, winning 2014 Kaggle competition



Galaxy Zoo 2: detailed morphological classifications for 304 122 galaxies from the Sloan Digital Sky Survey, *Willett, et al.* (2013)

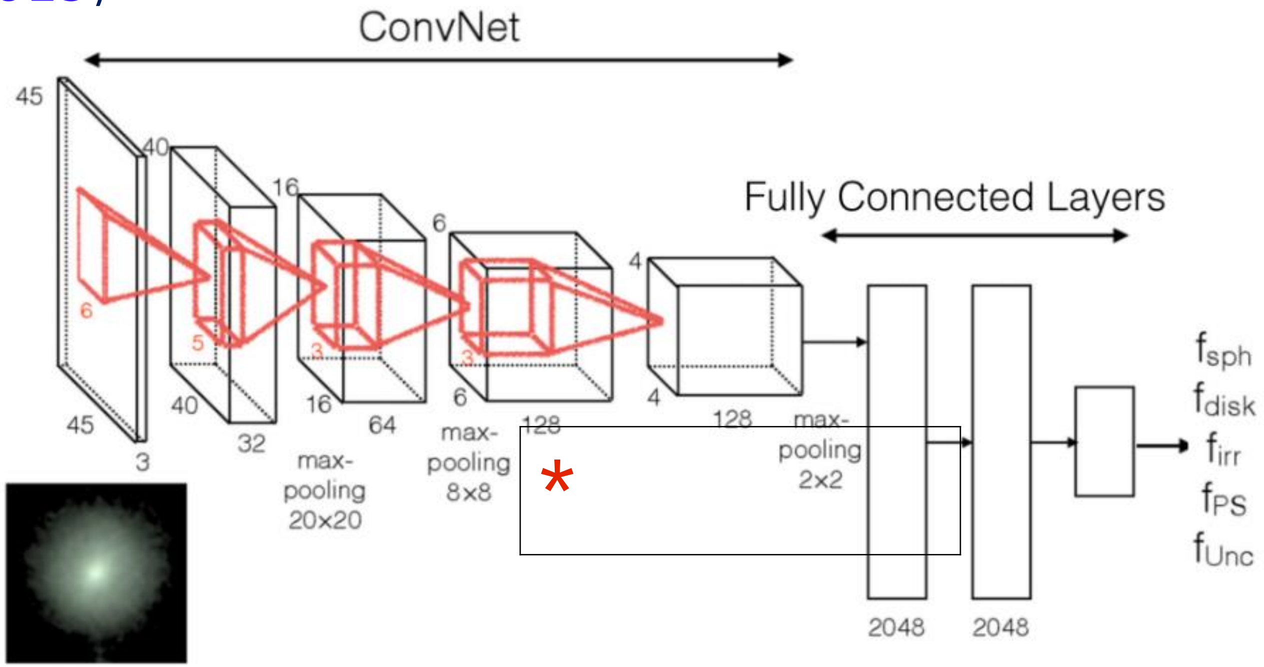
http://benanne.github.io/2014/04/05/galaxy-zoo.html



Rotation-invariant convolutional neural networks for galaxy morphology prediction, *Dieleman, et al.* (2015)

Huertas-Company et al. used CNN to classify **CANDELS*** galaxy images

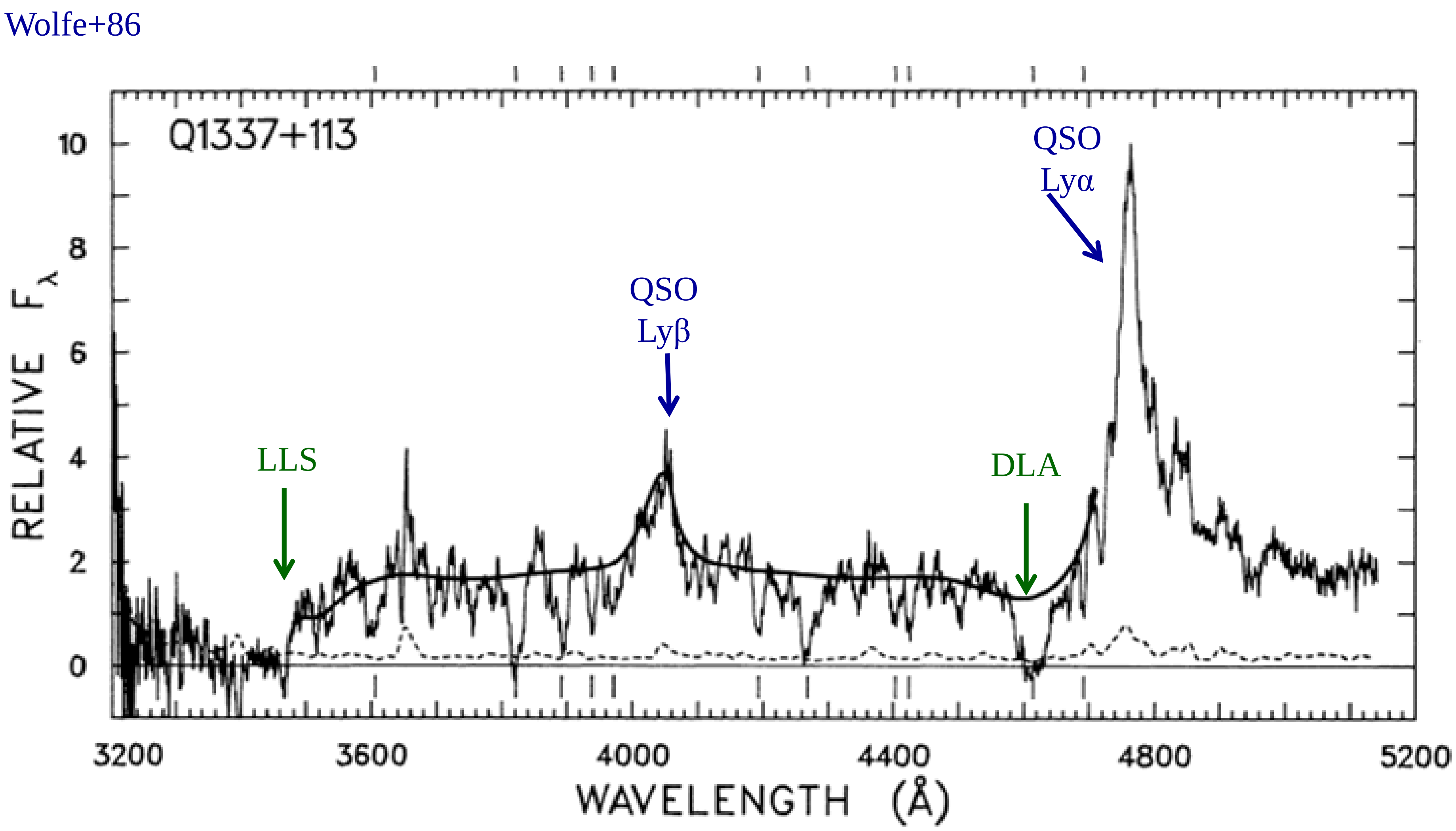
A Catalog of Visual-like Morphologies in the 5 CANDELS Fields Using Deep Learning, *Huertas-Company et al.* (2015)



Mass assembly and morphological transformations since z ~ 3 from CANDELS, *Huertas-Company et al.* (2016)

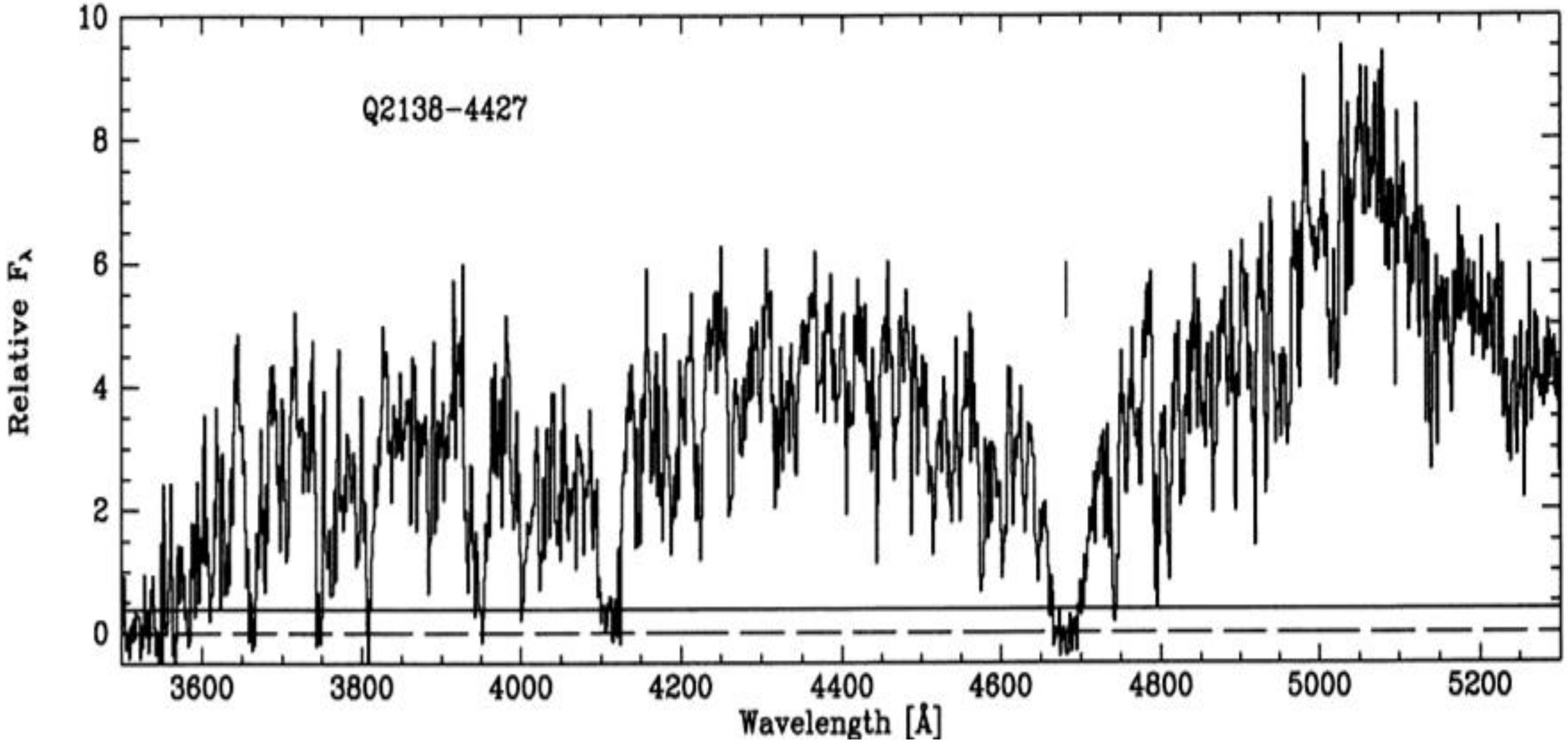
*CANDELS: Cosmic Assembly Near-infrared Deep Extragalactic Legacy Survey

The Damped Lyα Systems (DLAs)



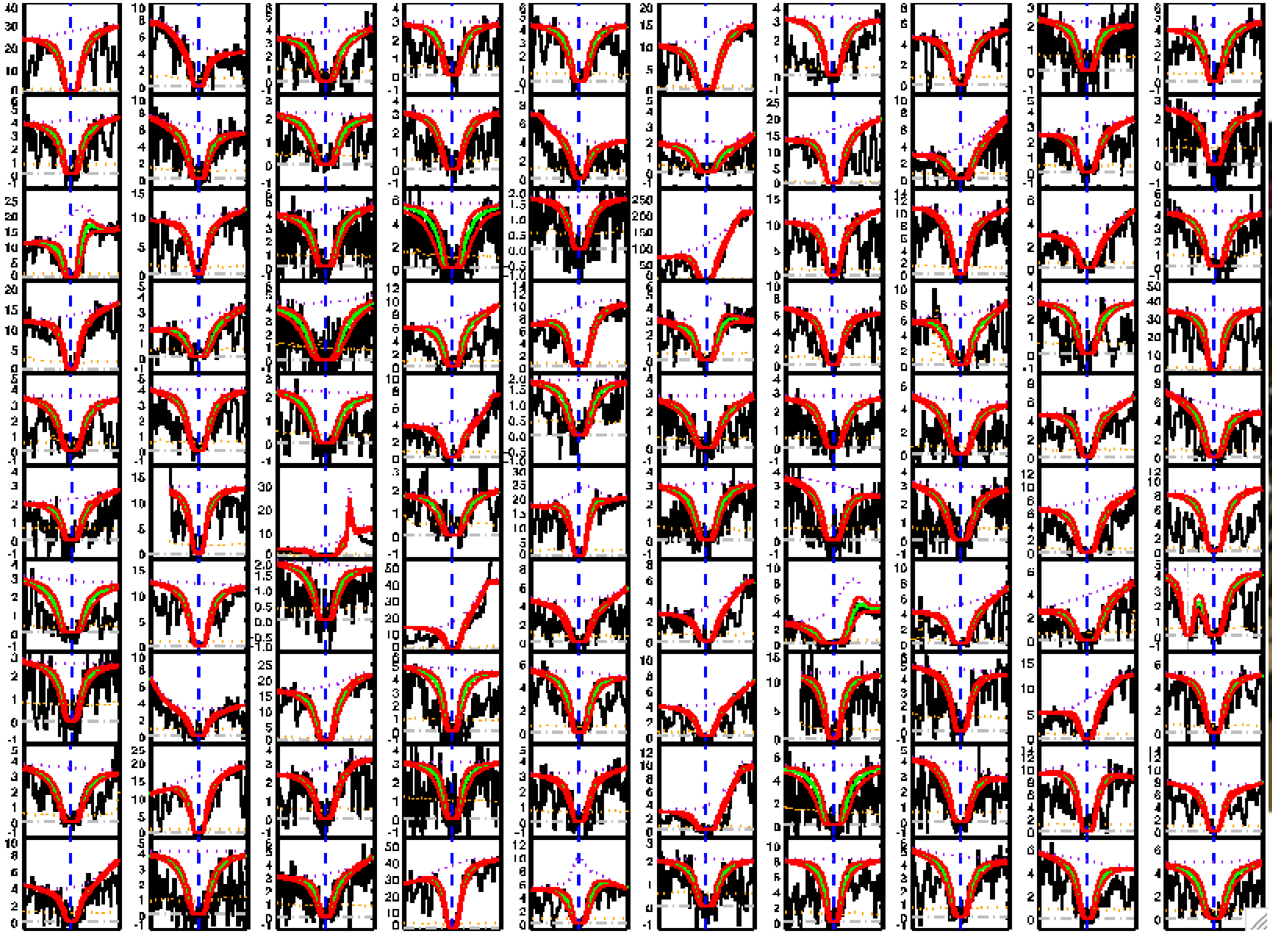
DLA Analysis (Old School)

Wolfe+95



Visual inspection (first error array!); by-eye N_{HI} fitting

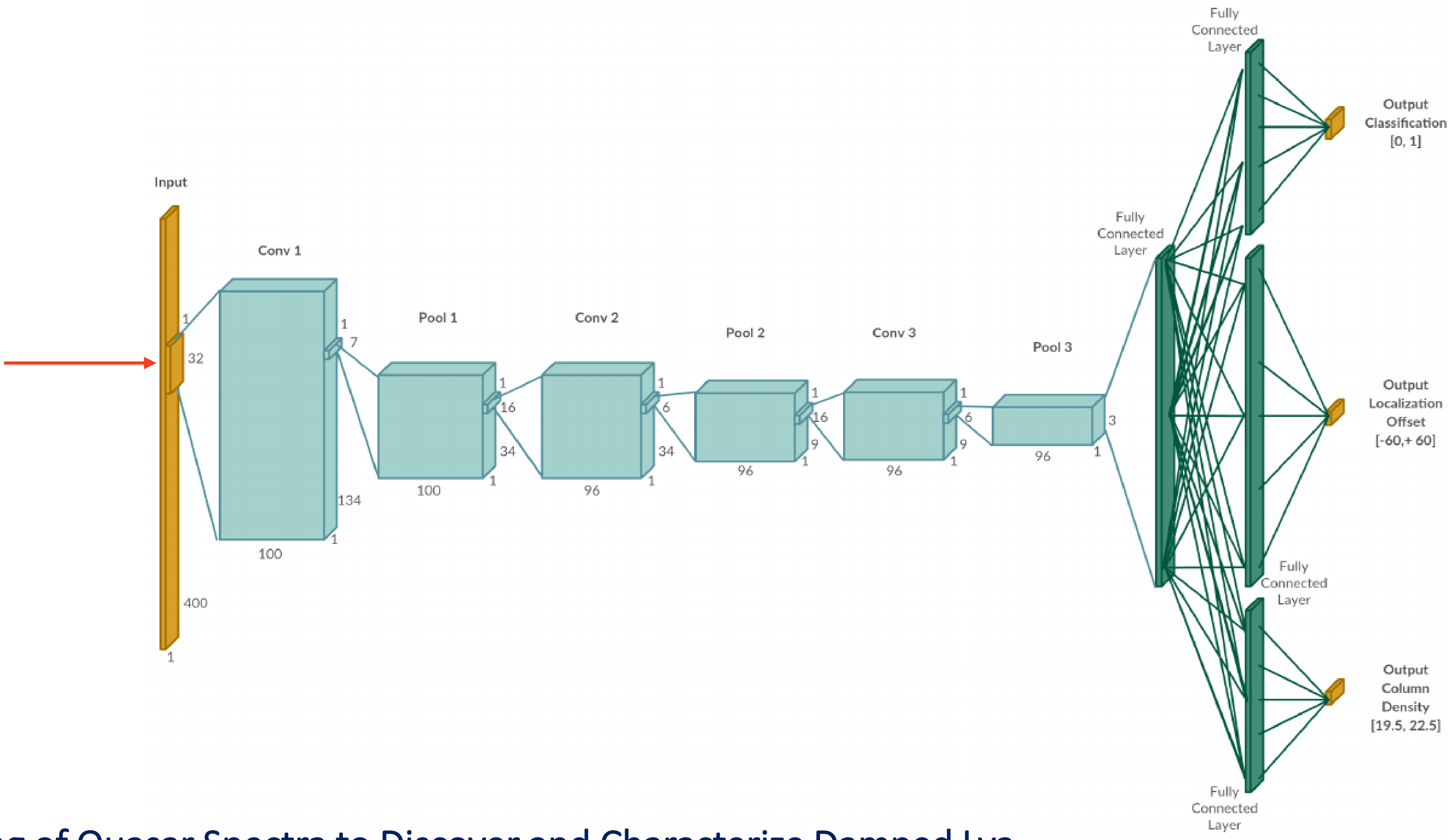
DLA Analysis (Early SDSS)



Visual inspection (first error array!); by-eye N_{HI} fitting

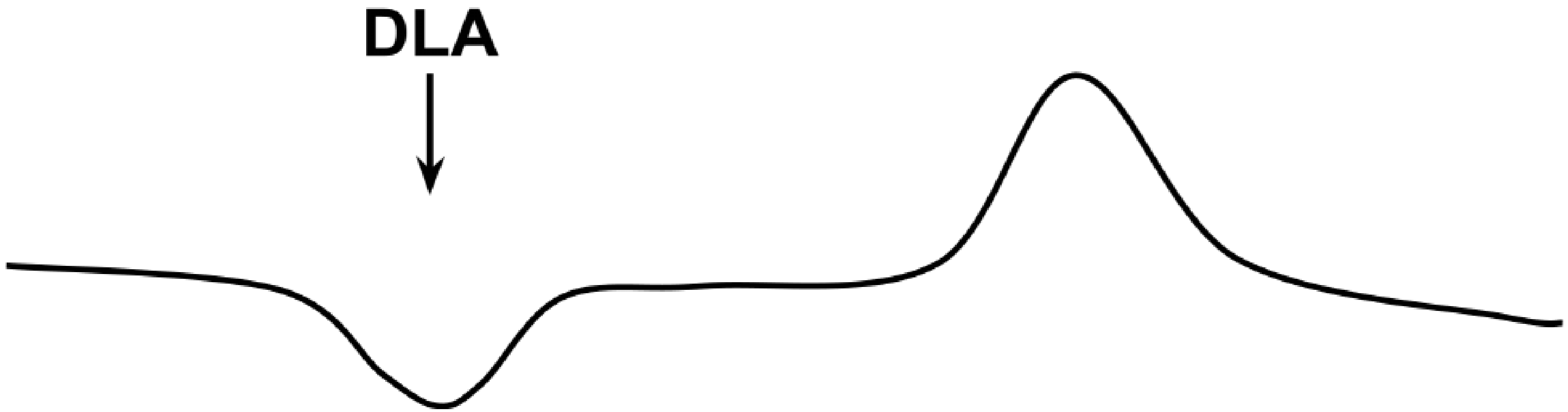
Deep Learning of DLAs

Cast the spectrum snippet as a 1D image



Deep Learning of Quasar Spectra to Discover and Characterize Damped Lya Systems, *Parks, Prochaska,* ***Dong****, & Cai* (2017)

CNN Labels and Learning



0	0	0	0	-60	...	-3	-2	-1	0	+1	+2	+3	...	+60	0	0	0	0	**Localization**
0	0	0	0	21.4	...	21.4	21.4	21.4	21.4	21.4	21.4	21.4	...	21.4	0	0	0	0	**NHI (Column Density)**
0	0	0	0	1	...	1	1	1	1	1	1	1	...	1	0	0	0	0	**Classification**

Three labels at each pixel:

1). Classification;
2). Localization (i.e., redshift);

3). HI column density

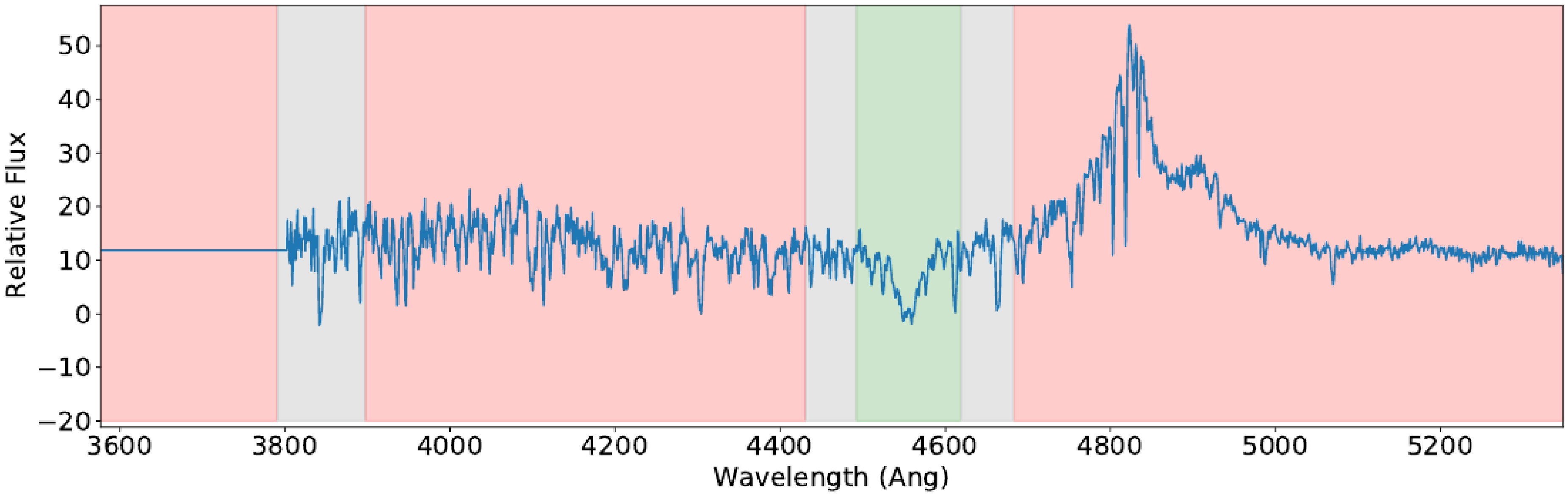
Multi-task Learning:
Combined loss function for all three labels

$$\mathcal{L}_c = -y_c \log(\hat{y}_c) - (1 - y_c) \log(1 - \hat{y}_c)$$

$$\mathcal{L}_o = (y_o - \hat{y}_o)^2$$

$$\mathcal{L}_h = \left(\frac{y_c}{y_c + \epsilon}\right)(y_h - \hat{y}_h)^2$$

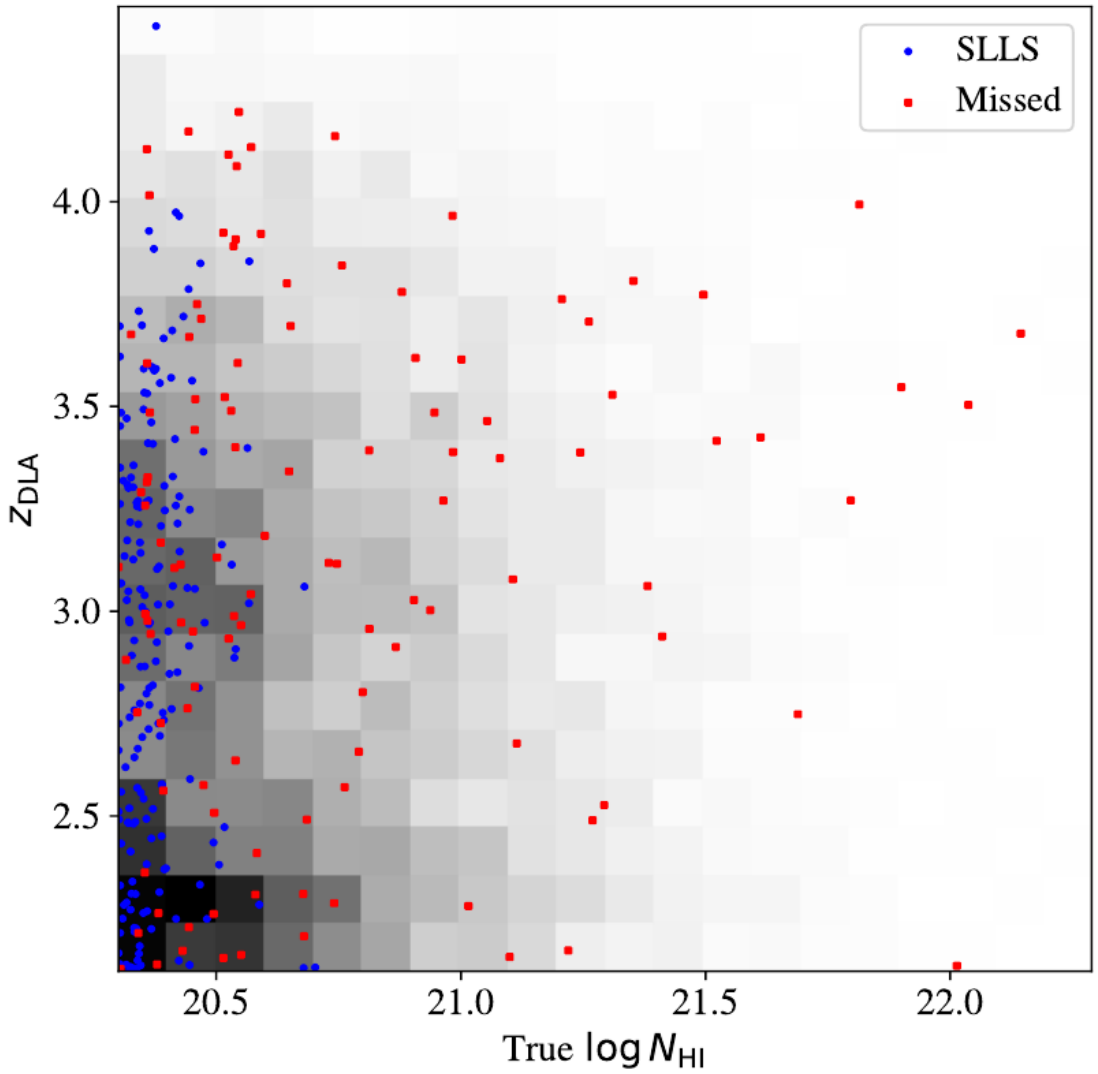
CNN Training



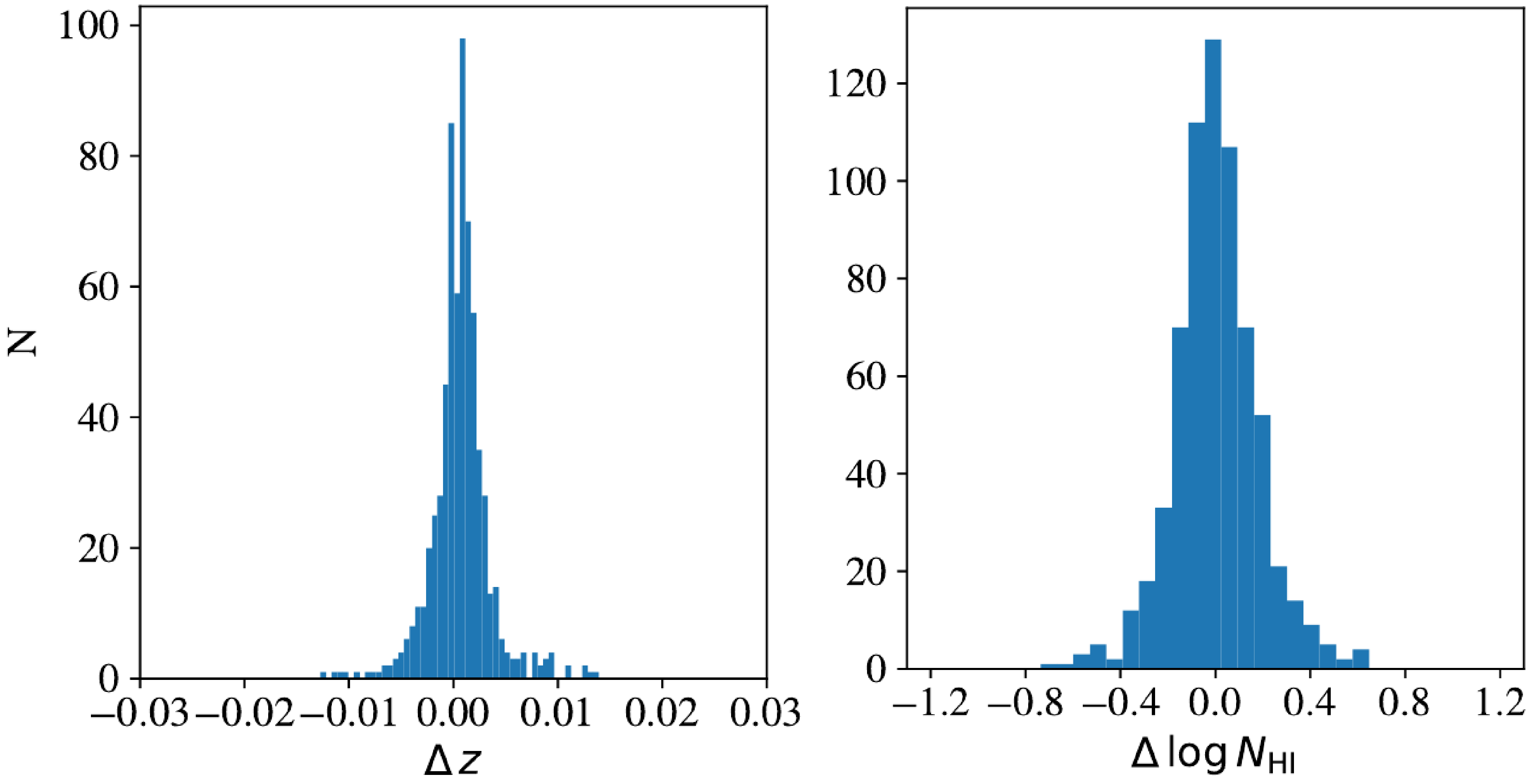
200,000 sightlines of DLAs injected into DLA-free quasar spectra from the SDSS-DR5.
By-hand addition of additional training sets: high N_{HI}, SLLS

CNN Validation

- 11,110 DLAs held back
  - Matched 99% of these to within |dz| < 0.015
- ~300 false negatives
  - Primarily SLLS
  - Overlapping DLAs
- 74 ‘false positives’
  - Some may be real
  - Overlapping DLAs

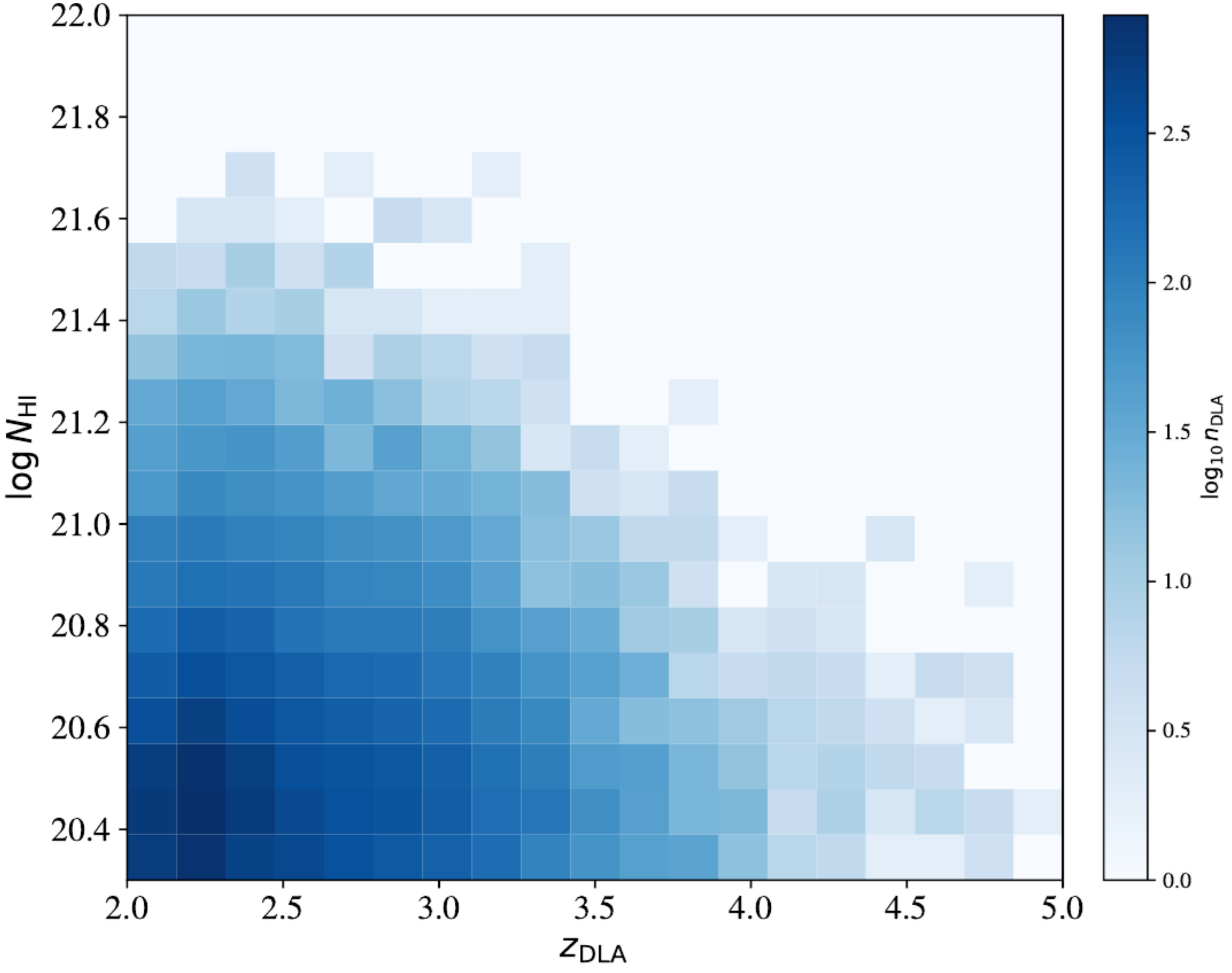


CNN Validation



Precise DLA measurements **without** Quantum Mechanics!!

DLA Results (CNN + BOSS)



found ~19,000 DLAs with z>2

Machine Learning in Astronomy

- Astronomy is rife with tasks demanding human labor
  - Source identification
  - Continuum fitting
  - Line identification
  - Etc.
- Machine Learning
  - Can perform many of these tasks
  - Auto-magically, repeatedly, better!