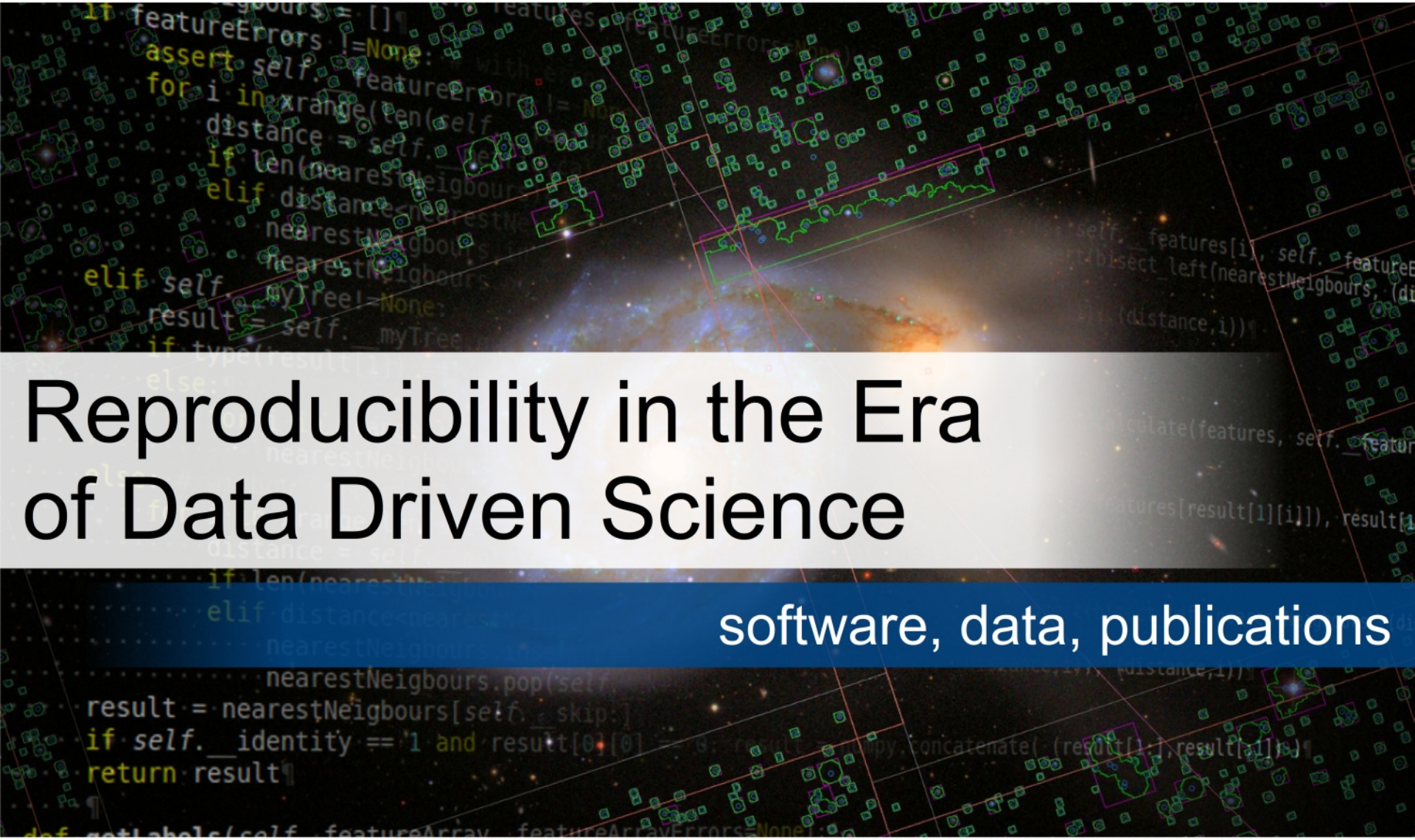
Heidelberg Institute for
Theoretical Studies

HITS



Reproducibility in the Era of Data Driven Science

software, data, publications

Reproducibility

multiple aspects:

- publications
- data / catalogs
- software / code / algorithms
- parameters / training- / test-data
- statistical methods

Reproducibility

multiple aspects:

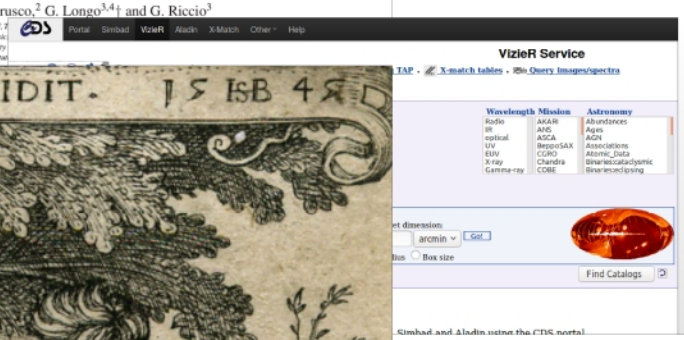
www.wikipedia.org

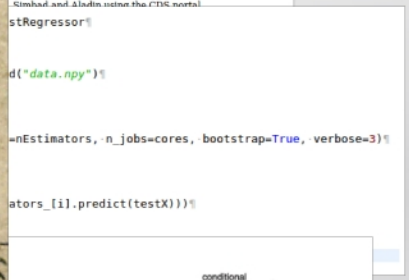

Monthly Notices
ROYAL ASTRONOMICAL SOCIETY

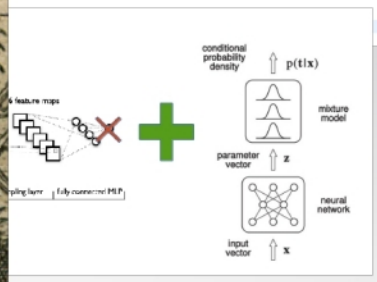
Mon. Not. R. Astron. Soc. 418, 2165–2195 (2011) doi:10.1111/j.1365-2966.2011.19416.x

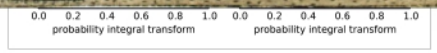
Astroinformatics of galaxies and quasars: a new general method for photometric redshifts estimation

O. Laurino,[1,2*] R. D'Abrusco,[2] G. Longo[3,4†] and G. Riccio[3]









Photometric Redshift Estimation

Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

Mon. Not. R. Astron. Soc. **418,** 2165–2195 (2011) doi:10.1111/j.1365-2966.2011.19416.x

Astroinformatics of galaxies and quasars: a new general method for photometric redshifts estimation

O. Laurino,[1,2★] R. D'Abrusco,[2] G. Longo[3,4]† and G. Riccio[3]

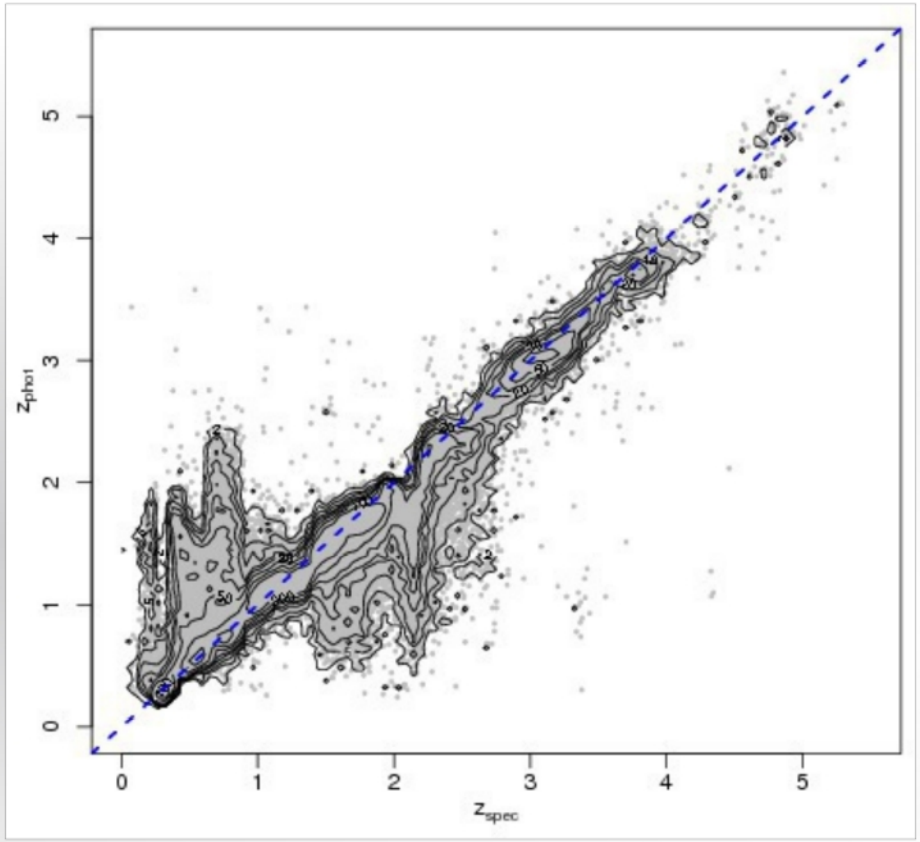
[1]*Astronomical Observatory of Trieste, INAF, Trieste 34143, Italy*
[2]*Harvard-Smithsonian Center for Astrophysics, Cambridge, MA 02138, USA*
[3]*Department of Physical Sciences, University of Naples, Naples 80126, Italy*
[4]*Department of Astronomy, California Institute of Technology, Pasadena, CA 90125, USA*

Accepted 2011 July 10. Received 2011 June 17; in original form 2011 March 9

ABSTRACT

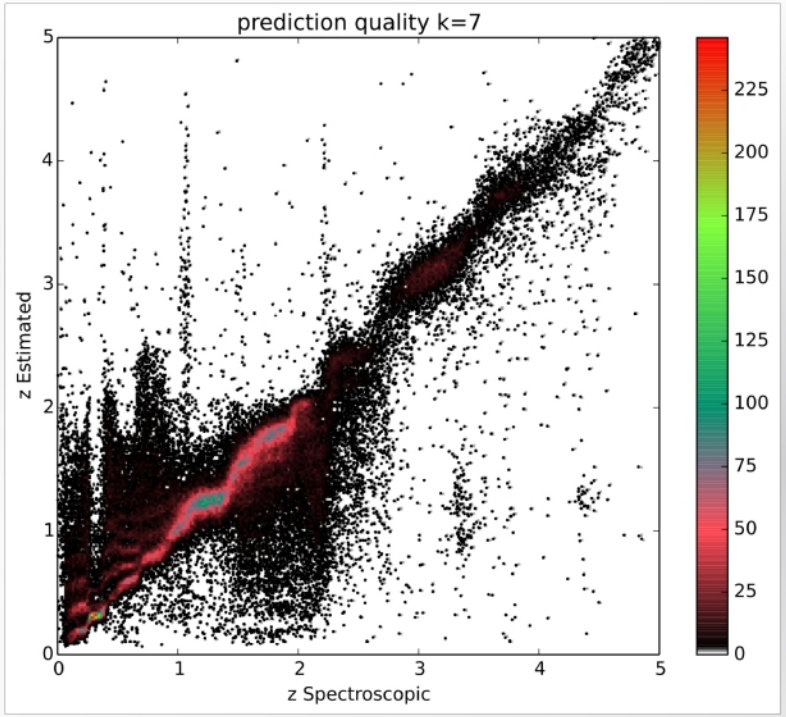
With the availability of the huge amounts of data produced by current and future large multi-band photometric surveys, photometric redshifts have become a crucial tool for extragalactic astronomy and cosmology. In this paper we present a novel method, called Weak Gated Experts (WGE), which allows us to derive photometric redshifts through a combination of data mining techniques. The WGE, like many other machine learning techniques, is based on the exploitation of a spectroscopic knowledge base composed by sources for which a spectroscopic value of the redshift is available. This method achieves a variance $\sigma^2(\Delta z) = 2.3 \times 10^{-4}$ [$\sigma^2(\Delta z) = 0.08$, where $\Delta z = z_{\rm phot} - z_{\rm spec}$] for the reconstruction of the photometric redshifts

Photometric Redshift Estimation



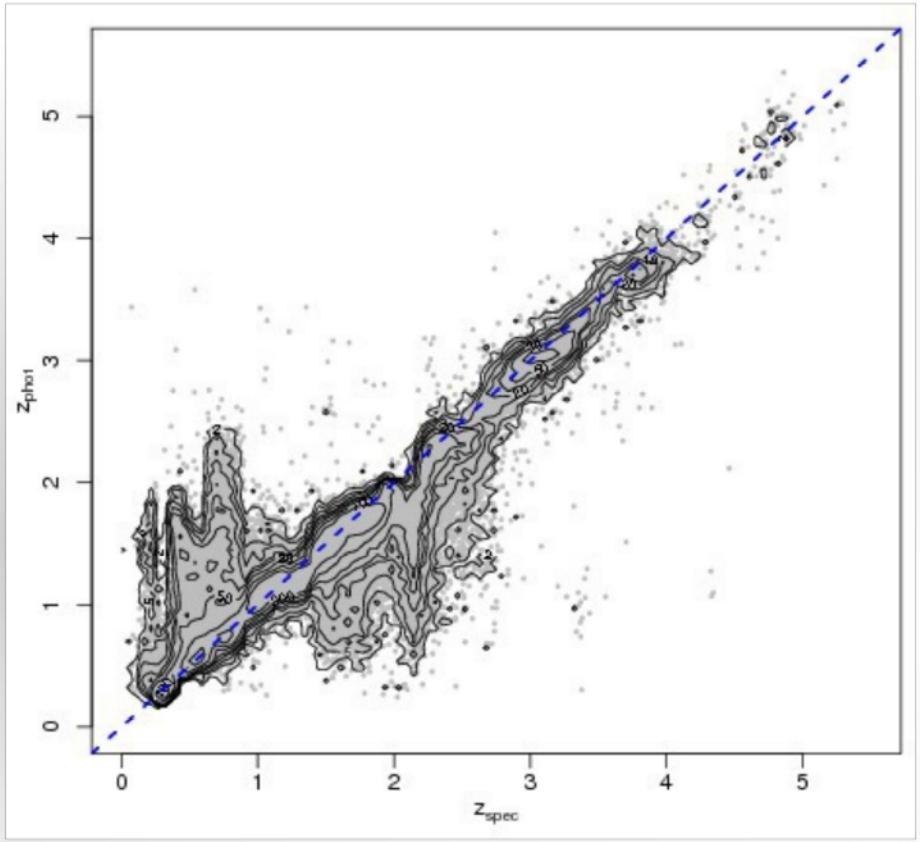
Laurino et al. 2011

$$RMSE(\Delta z_{norm}) = 0.19$$
$$MAD(\Delta z_{norm}) = 0.041$$



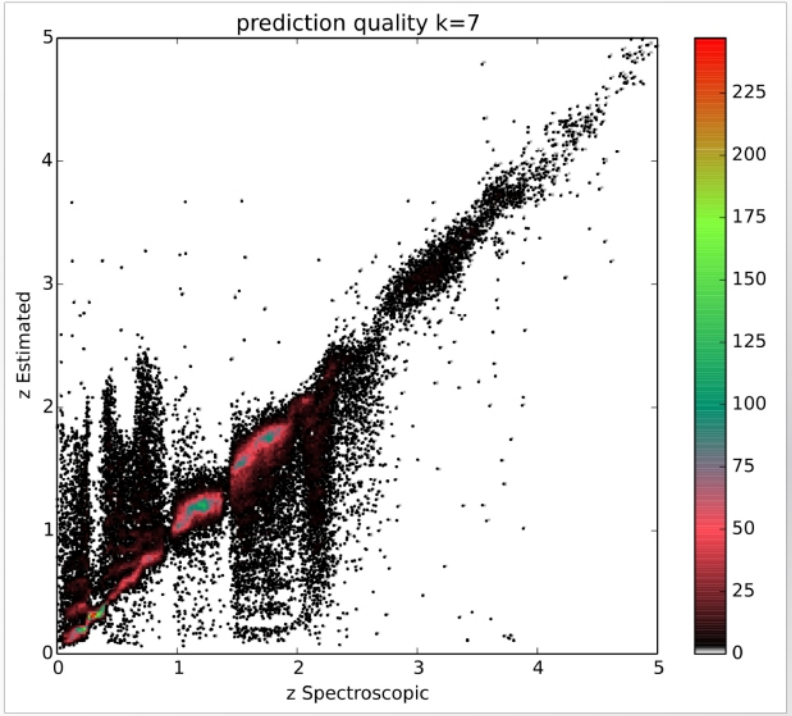
$$RMSE(\Delta z_{norm}) = 0.25$$
$$MAD(\Delta z_{norm}) = 0.048$$

Photometric Redshift Estimation



Laurino et al. 2011

$$RMSE(\Delta z_{norm}) = 0.19$$
$$MAD(\Delta z_{norm}) = 0.041$$



$$RMSE(\Delta z_{norm}) = 0.22$$
$$MAD(\Delta z_{norm}) = 0.038$$

Publishing Training- / Test-data

more than just:

- SDSS DR7 etc.
- VizieR ID
- select statement

required:

- catalogs
- labels / annotations
- data-bases
- reference data-sets
- scripts / pre-processing

```
SELECT TOP 200000
p.objid,p.ra,p.dec,p.u,p.g,p.r,p.i,p.z,
p.psfMag_u, p.psfMag_g, p.psfMag_r, p.psfMag_i,
p.psfMag_z, p.modelMag_u, p.modelMag_g, p.modelMag_r,
p.modelMag_i, p.modelMag_z, s.specobjid, s.class,
s.z AS redshift
INTO mydb.DR9_galaxies_with_modMag
FROM PhotoObj AS p
JOIN SpecObj AS s ON s.bestobjid = p.objid
WHERE s.z BETWEEN 0 AND 6.0
AND s.class = 'GALAXY'
ORDER BY NEWID()
```

Publishing and Sharing Code

HITS

simple code / simple example !?

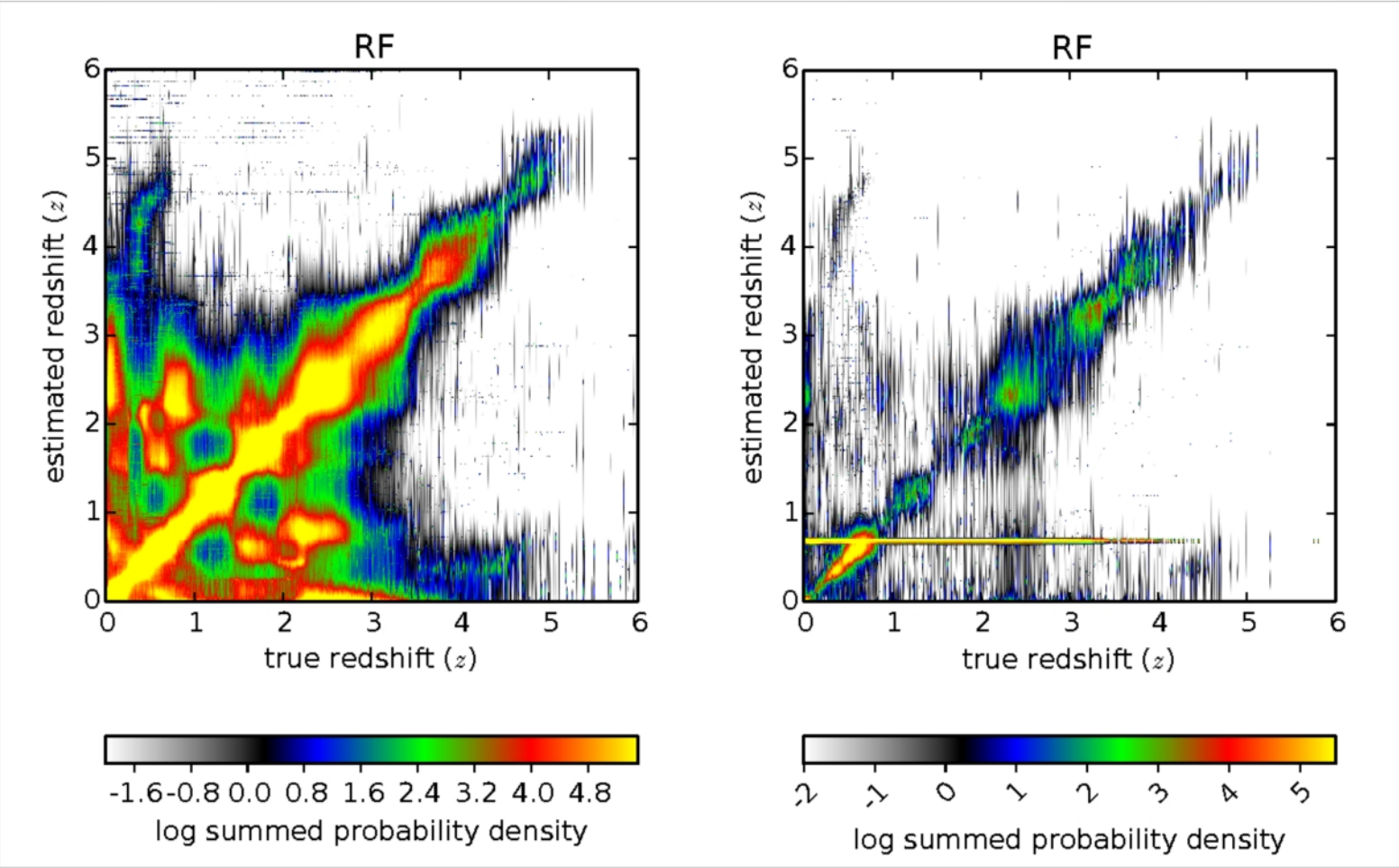
- quantile regression forest for probabilistic predictions

```
1  from sklearn.ensemble import RandomForestRegressor
2  from sklearn.mixture import GMM
3  import numpy
4
5  trainX, trainY, testX,testY = numpy.load("data.npy")
6
7  nEstimators = 256
8  cores = 8
9
10 rf = RandomForestRegressor(n_estimators=nEstimators, n_jobs=cores, bootstrap=True, verbose=3)
11 rf.fit(trainX, trainY)
12
13 results = []
14 for i in range(len(rf.estimators_)):
15     results.append(numpy.array(rf.estimators_[i].predict(testX)))
16 results = numpy.array(results).T
17
18 for i in range (len(testY)):
19     myModel = GMM(5, min_covar=0.00001).fit(results[i])
```

Publishing and Sharing Code

HITS

same code on 8x core and 120x core machine



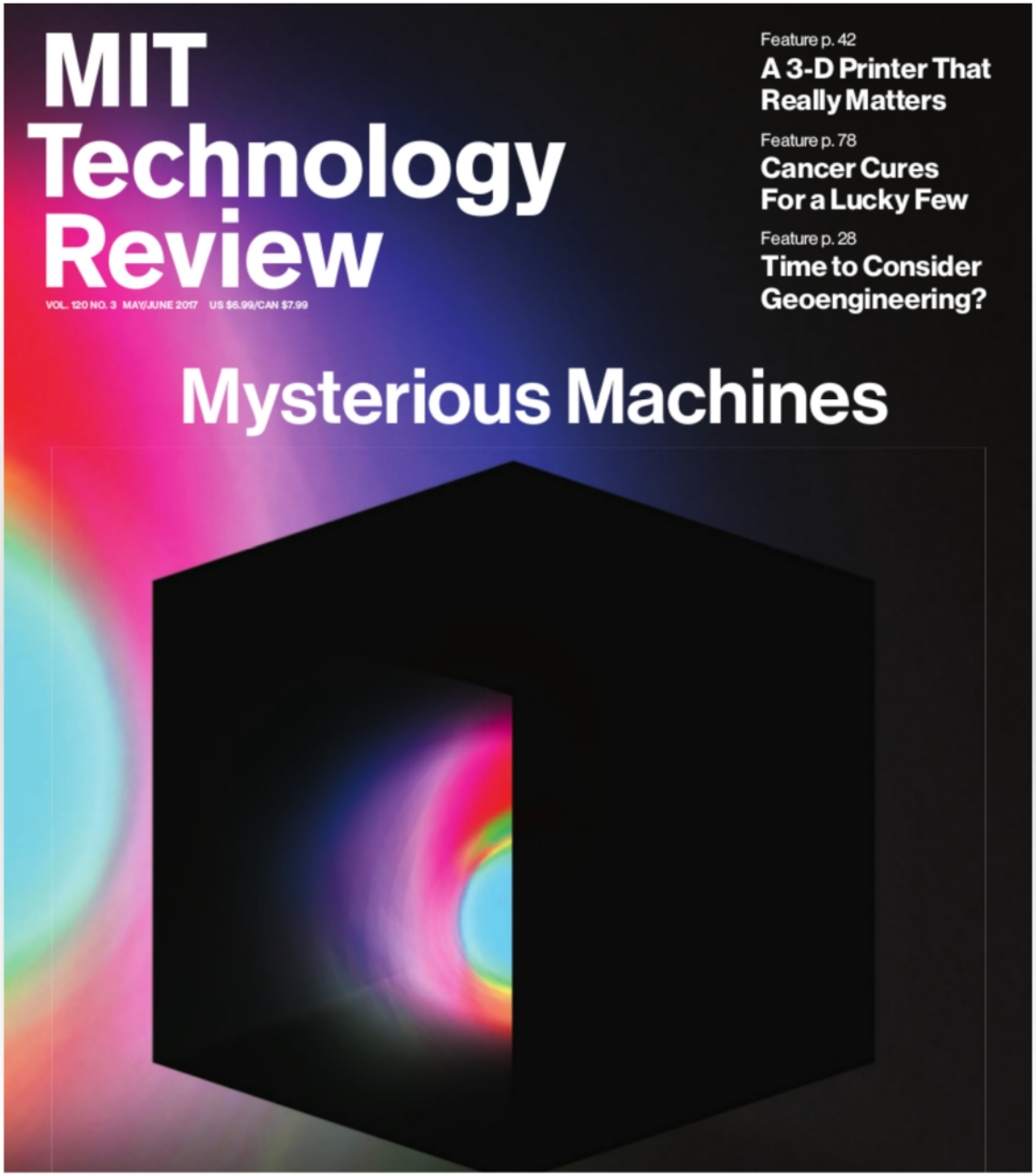
Publishing and Sharing Code

it is mandatory to publish code ...

… but how to publish

- environments
  - hardware
  - software / libraries
  - container
- compiler / interpreter
- test-data to verify the code is running as expected

… how to preserve and publish functionality ?

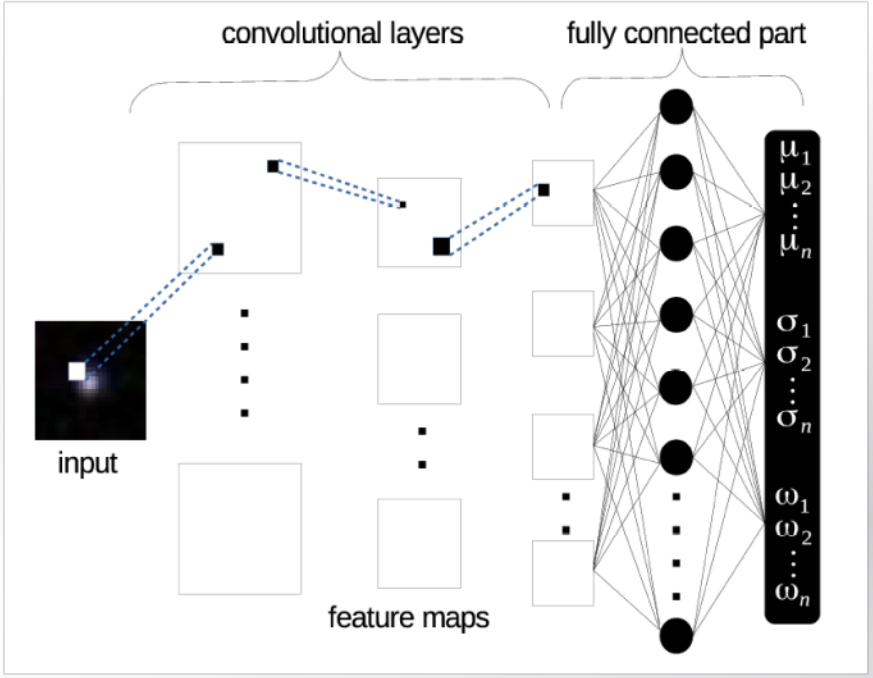

MIT Technology Review
VOL. 120 NO. 3 MAY/JUNE 2017 US $6.99/CAN $7.99
Feature p. 42
A 3-D Printer That Really Matters
Feature p. 78
Cancer Cures For a Lucky Few
Feature p. 28
Time to Consider Geoengineering?
Mysterious Machines

HITS

Machine Learning

publish all:

- architecture / model
- hyper-parameters



deep-learning:

- weights / biases
- training-data / data augmentation

online-learning / streaming algorithms:

- how to preserve this?

Statistical Methods

John P. A. Ioannidis: “Why Most Published Research Findings Are False”

NCBI Resources How To Sign in to NCBI

PMC US National Library of Medicine National Institutes of Health — PMC — Search — Advanced Journal list Help

Journal List > PLoS Med > v.2(8); 2005 Aug > PMC1182327

PLOS | MEDICINE A Peer-Reviewed, Open Access Journal

View this Article | Submit to PLOS | Get E-Mail Alerts | Contact Us

PLoS Med. 2005 Aug; 2(8): e124. PMCID: PMC1182327
Published online 2005 Aug 30. doi: 10.1371/journal.pmed.0020124

Why Most Published Research Findings Are False

John P. A. Ioannidis

Author information ► Copyright and License information ►

See "Minimizing Mistakes and Embracing Uncertainty" , e272.
See "Power, Reliability, and Heterogeneous Results" in volume 2, e386.
See "The Clinical Interpretation of Research" in volume 2, e395.
See "Author's Reply" in volume 2, e398.
See "Truth, Probability, and Frameworks" in volume 2, e361.
See "Why Most Published Research Findings Are False: Problems in the Analysis" in volume 4, e168.
See "Why Most Published Research Findings Are False: Author's Reply to Goodman and Greenland" in volume 4, e215.
See "Why Current Publication Practices May Distort Science" in volume 5, e201.

This article has been cited by other articles in PMC.

Summary Go to:

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

PLoS Med PLoS Med

Formats: Article | PubReader | ePub (beta) | PDF (250K) | Citation

Share Facebook Twitter Google+

Save items Add to Favorites

Similar articles in PubMed

Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-n [Health Technol Assess. 2001]

Estimating a summarized odds ratio whilst eliminating publication bias in meta-analysis. [Jpn J Clin Oncol. 1992]

Effect of formal statistical significance on the credibility of observational associations. [Am J Epidemiol. 2008]

Concepts in sample size determination. [Indian J Dent Res. 2012]

Power dressing and meta-analysis: incorporating power analysis into meta-analysis. [J Adv Nurs. 2002]

See reviews... See all...

Links Cited in Books | PubMed | Taxonomy

Recent Activity Turn Off Clear

Why Most Published Research Findings Are False

Statistical Methods

proper scores / scoring rules

dealing with uncertainties

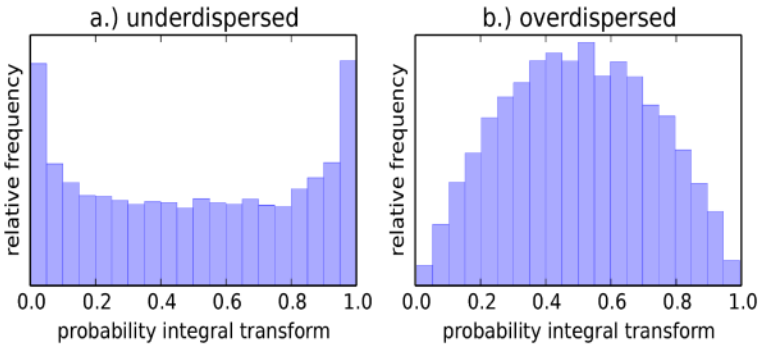
- continuous ranked probability score

$$CRPS = \frac{1}{N}\sum_{t=1}^{N} crps(CDF_t, z_t),$$

$$\text{with } crps(CDF_t, z_t) = \int_{-\infty}^{+\infty} [CDF_t(z) - CDF_{z_t}(z)]^2 \, dz$$

- probability integral transform

preserve data for different kinds of analysis !



Conclusion

publications

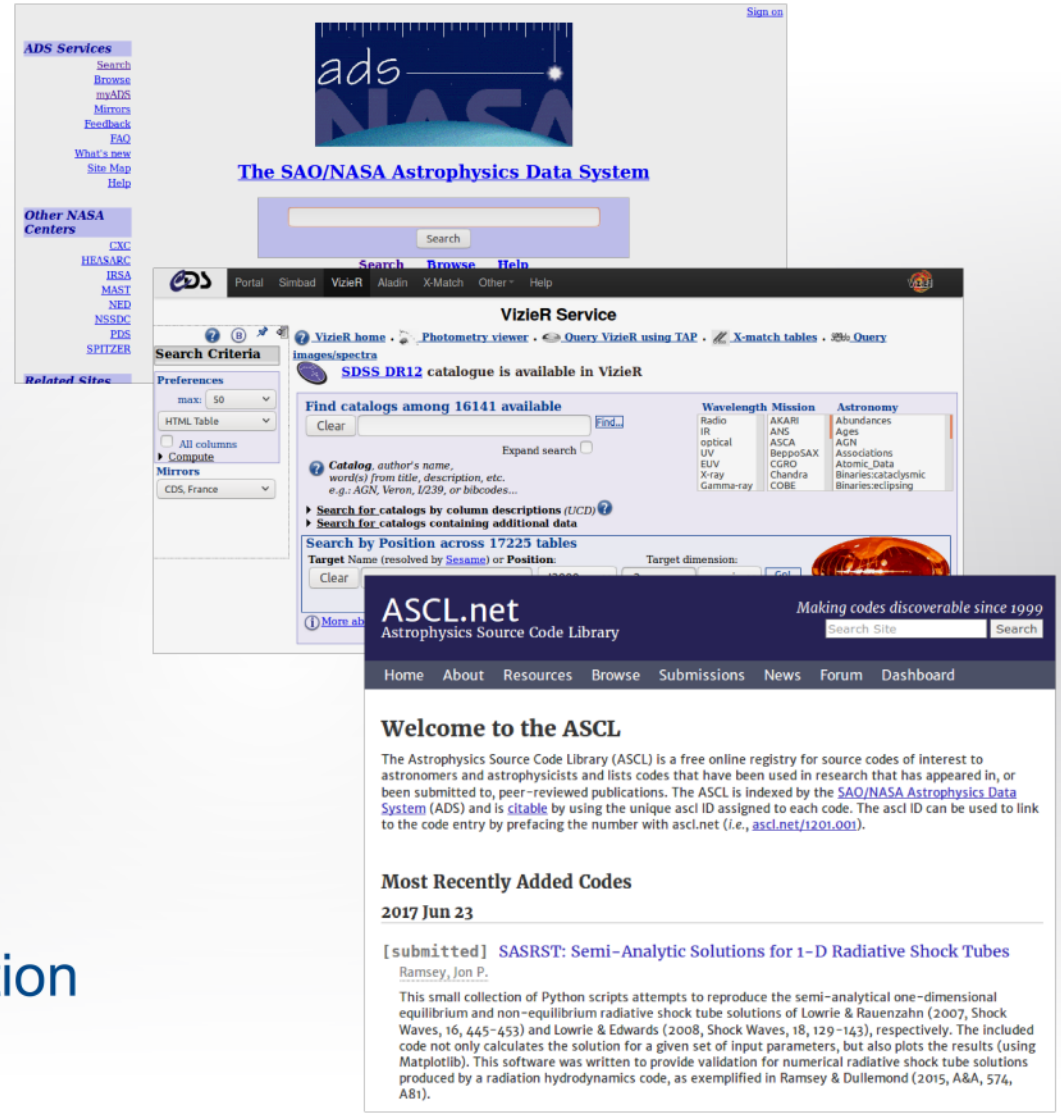
- open access

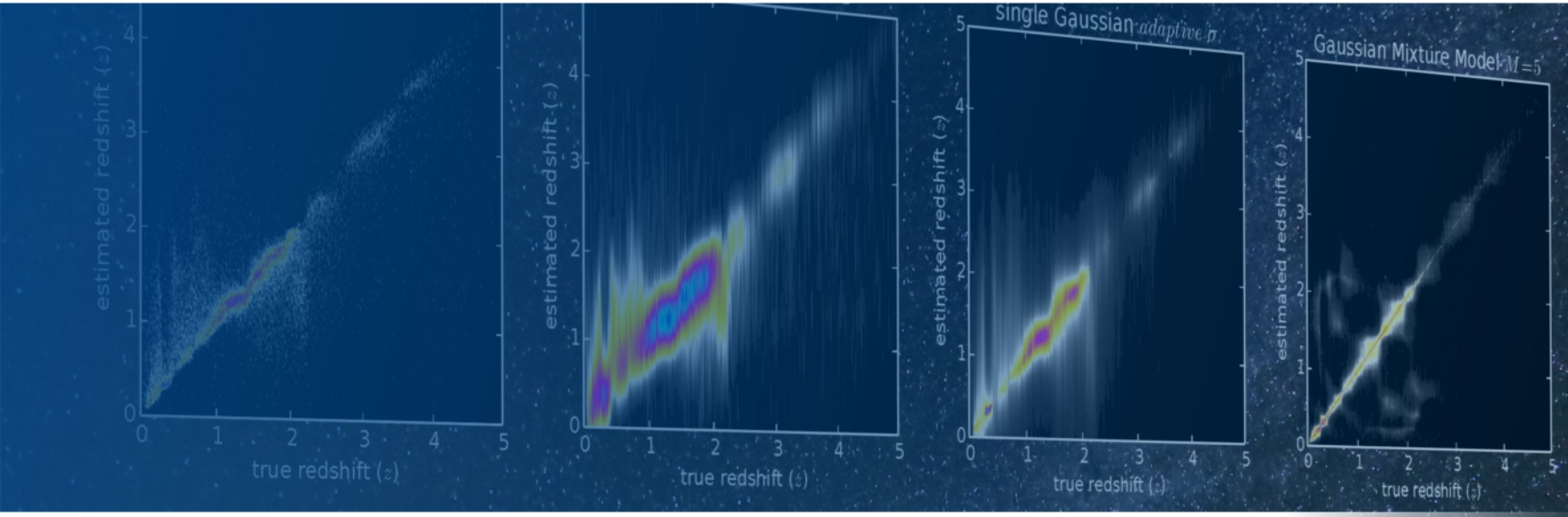
data

- raw-data / catalogs
- training- / test-data
- reference data-sets
- detailed results

software

- code / repositories
- parameter / configuration
- environment





thanks for your attention!

AiN

kai.polsterer@h-its.org / twitter:@Astroinformatix