ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

# Photo-z's PDF I/O, DB, DES (Lessons learned ... so far)

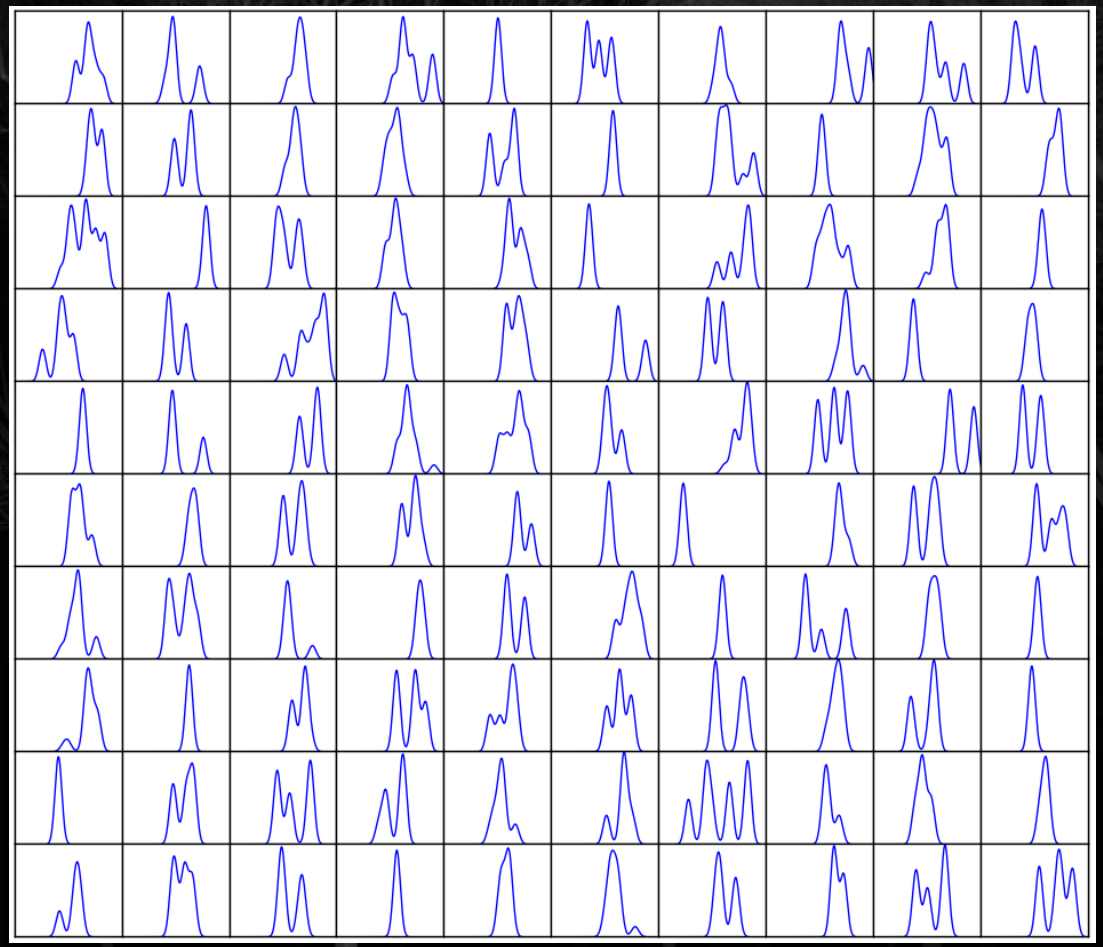
Matías Carrasco Kind
NCSA/Department of Astronomy
University of Illinois at Urbana-Champaign

LSST_DESC @ Argonne
October 29$^{\text{th}}$, 2015

# Photo-$z$ PDF in DES

- Photo-z PDF becoming important
- 10 DES papers using photo-z (WL, 2pt WL, 2pt LSS, DES X CMB, Systematics, redmagic, etc)
- Asorey et al 2015 in prep. includes PDF in clustering analysis.
- Storage and distribution still an issue
- Compression techniques

# Photo-$z$ PDF representation and storage in DES DB



# Photo-$z$ PDF storage: Strategies

Single Gaussian fit

Multi-Gaussan fit

Monte Carlo sampling

Sparse representation techniques

Reduce number of points while increasing accuracy

# Photo-$z$ PDF storage: Strategies

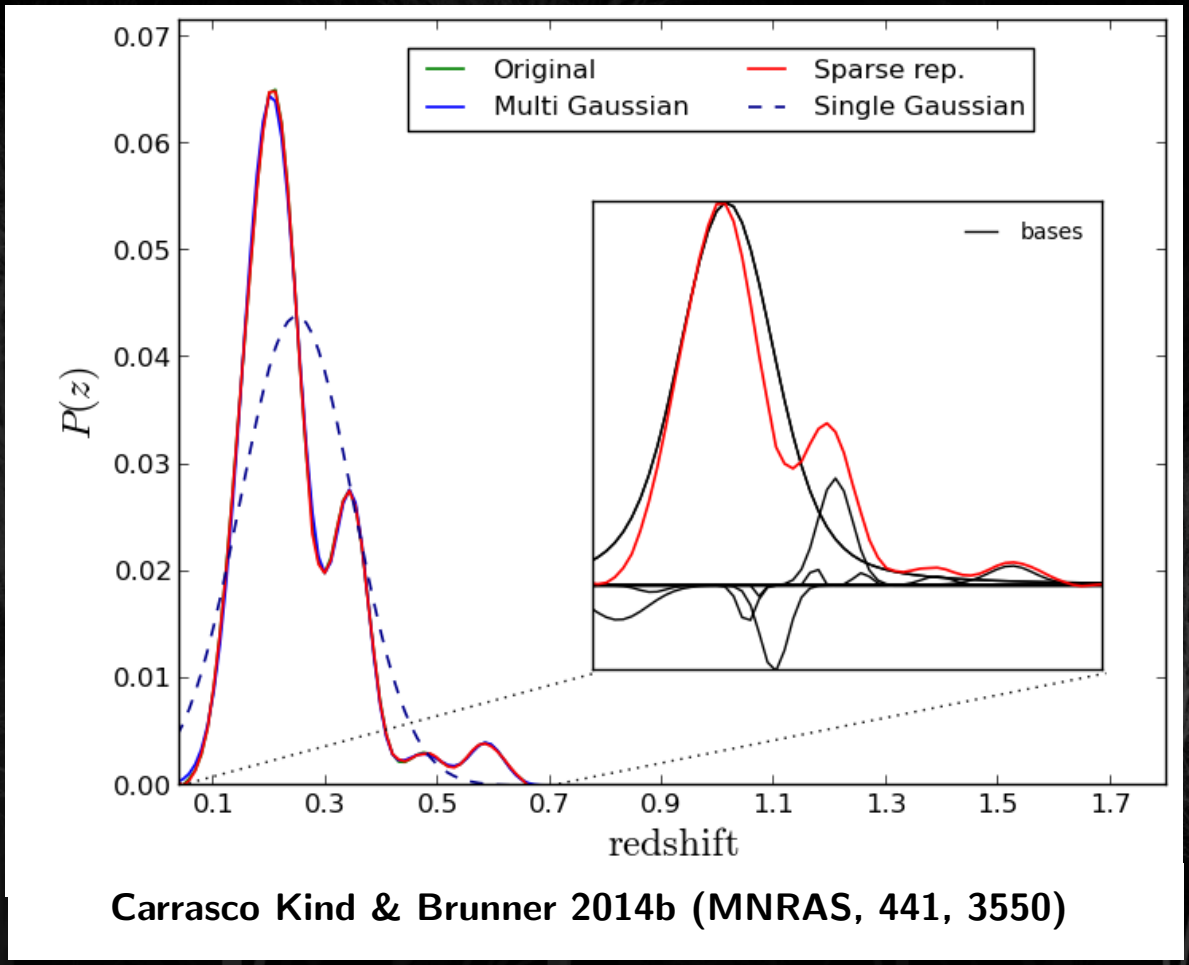
Single Gaussian fit

Multi-Gaussan fit

Monte Carlo sampling

Sparse representation techniques

Reduce number of points while increasing accuracy




**Carrasco Kind & Brunner 2014b (MNRAS, 441, 3550)**

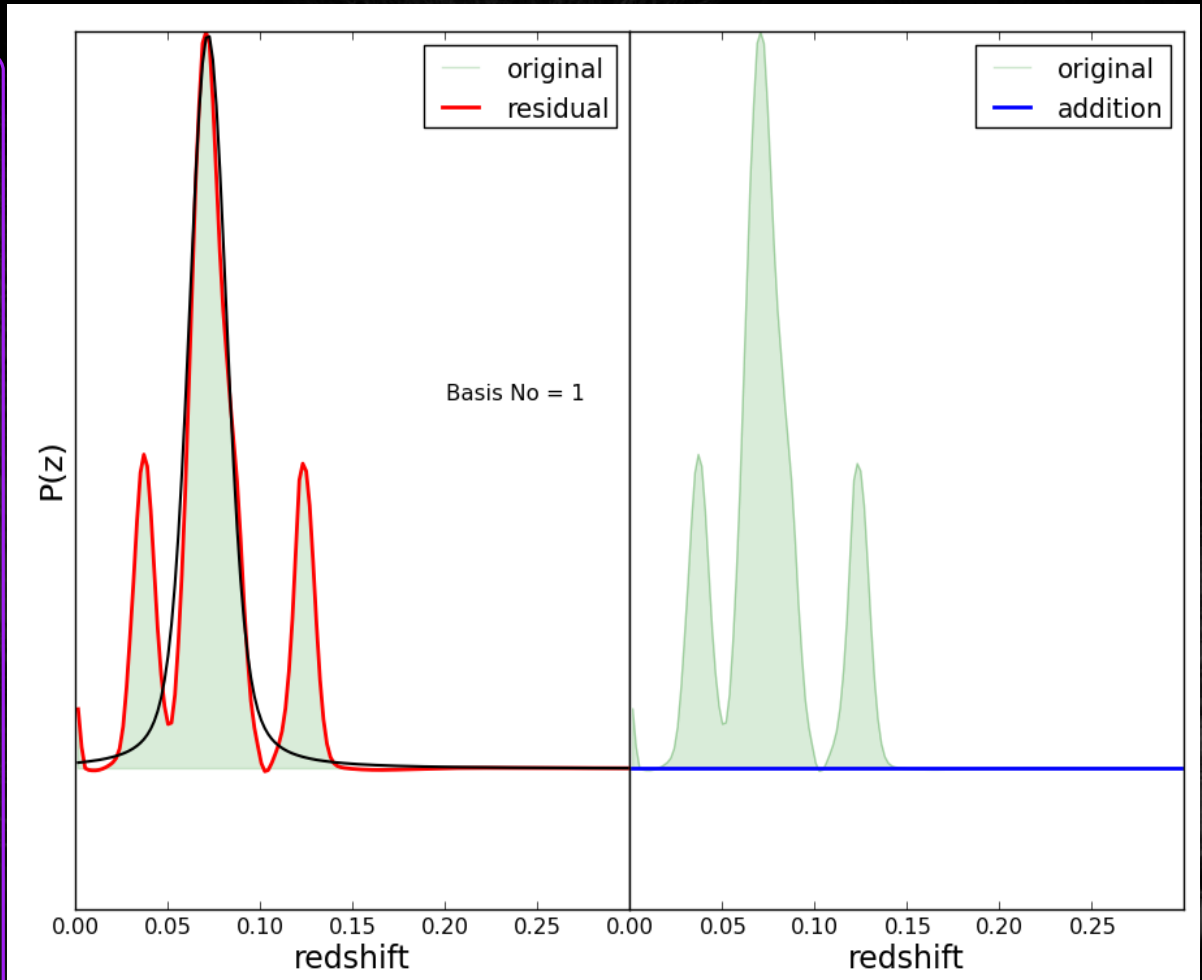
# Photo-$z$ PDF storage: Sparse representation

Use Gaussian and Voigt profiles as bases, need $N^2_{\rm original}$ bases

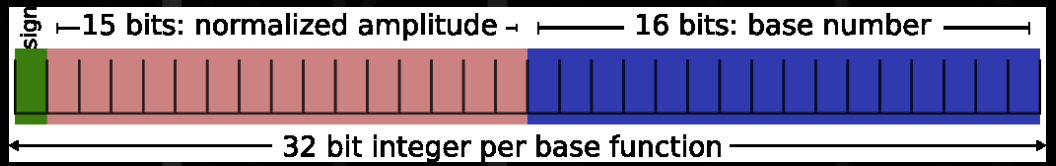
With only 10-20 bases achieve 99.9 % accuracy

Use 32-bits integer per basis, compression

Store Multiple PDFs




Carrasco Kind & Brunner 2014b (MNRAS, 441, 3550)



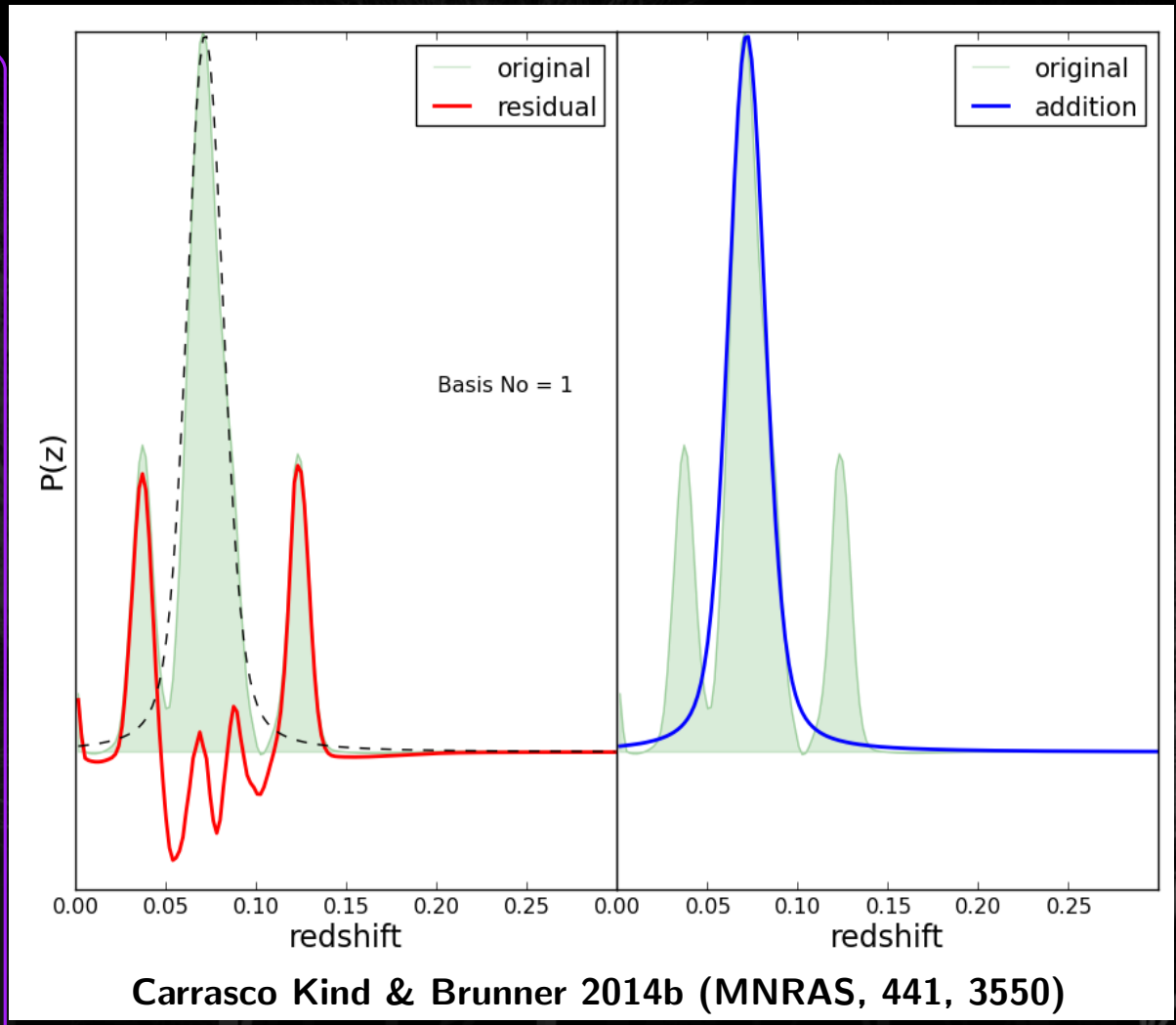
# Photo-$z$ PDF storage: Sparse representation

Use Gaussian and Voigt profiles as bases, need $N^2_{\rm original}$ bases

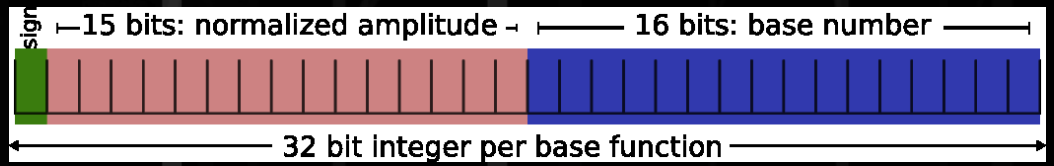
With only 10-20 bases achieve 99.9 % accuracy

Use 32-bits integer per basis, compression

Store Multiple PDFs




Carrasco Kind & Brunner 2014b (MNRAS, 441, 3550)



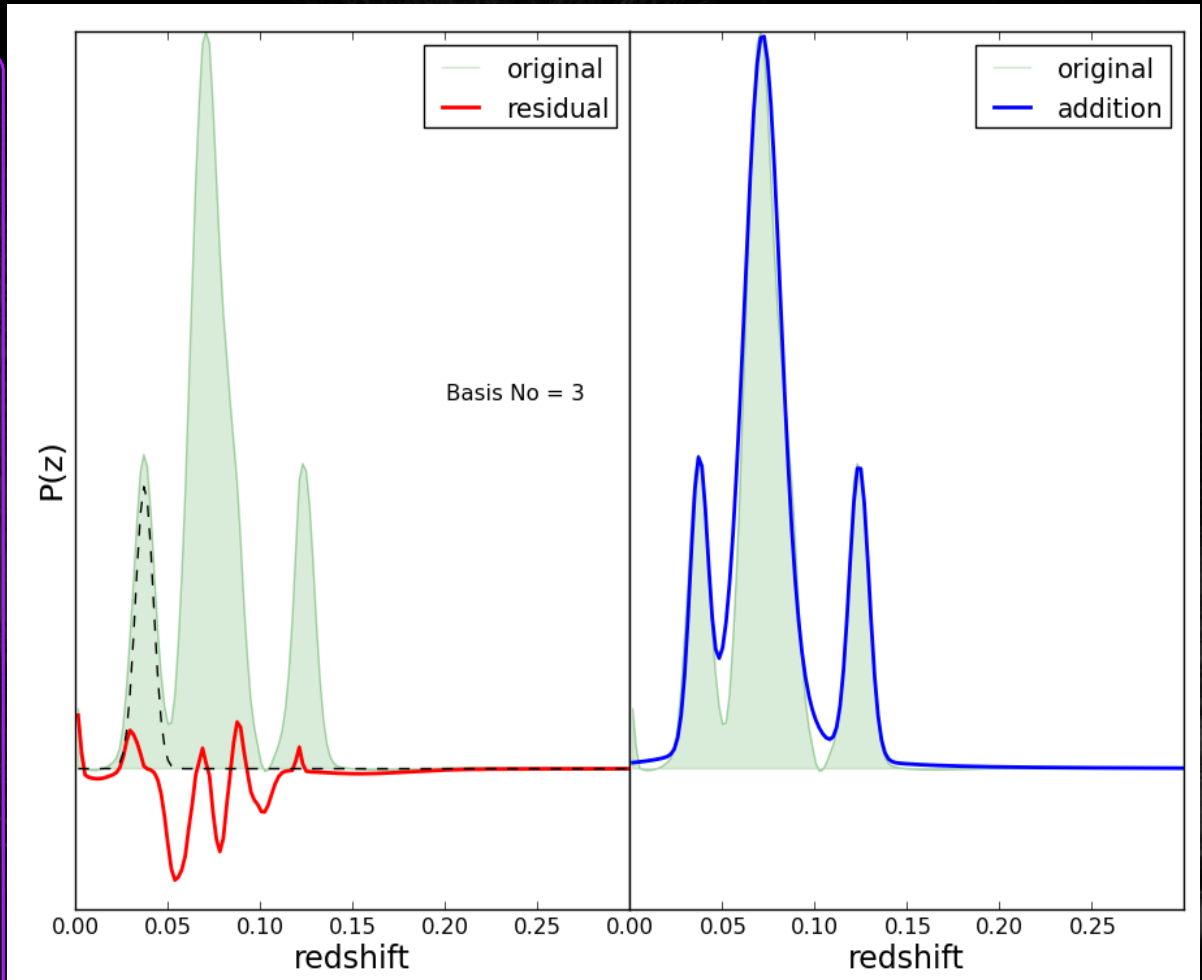
# Photo-$z$ PDF storage: Sparse representation

Use Gaussian and Voigt profiles as bases, need $N^2_{\rm original}$ bases

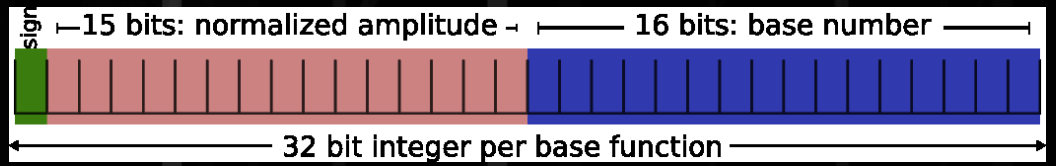
With only 10-20 bases achieve 99.9 % accuracy

Use 32-bits integer per basis, compression

Store Multiple PDFs




Carrasco Kind & Brunner 2014b (MNRAS, 441, 3550)



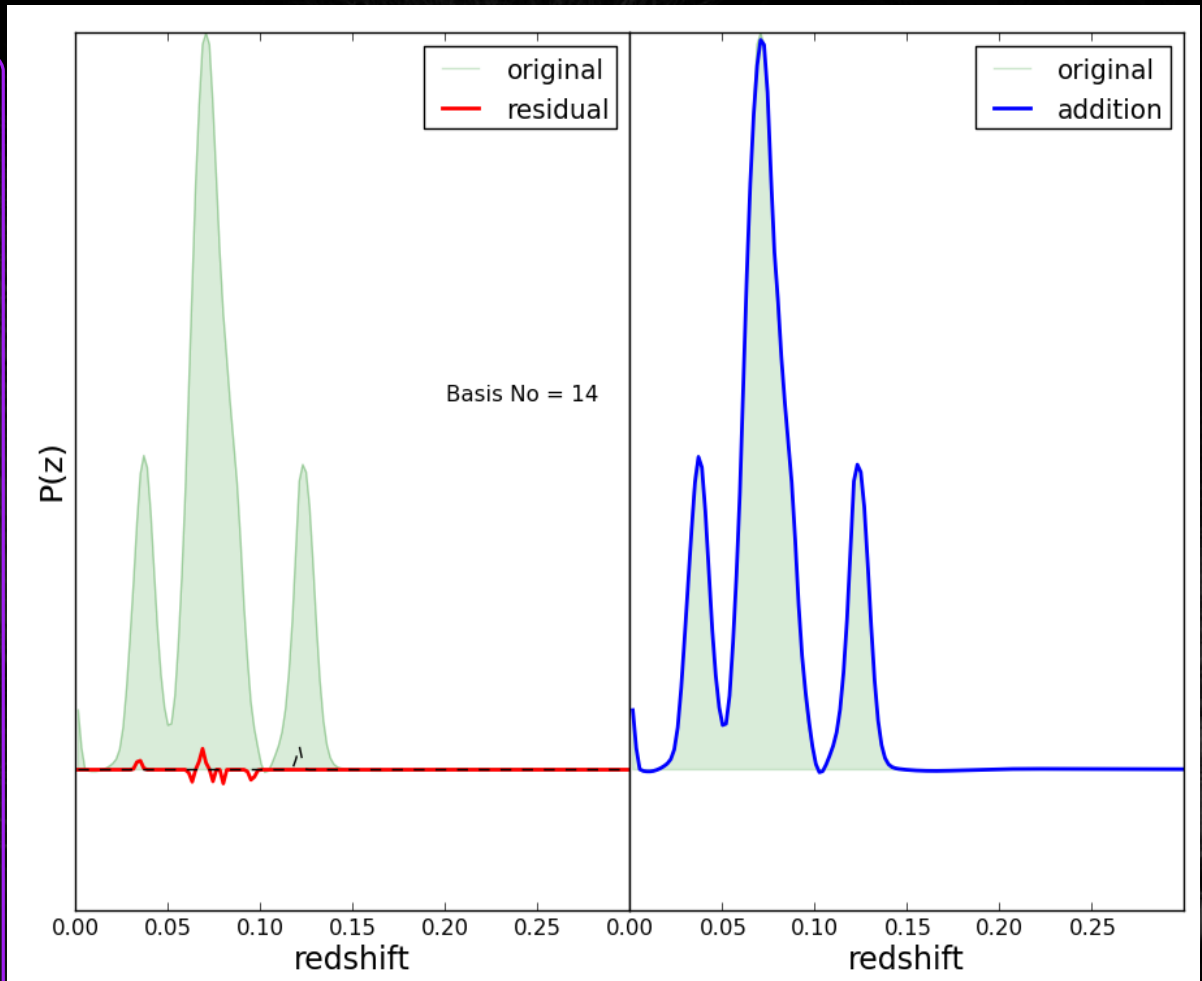
# Photo-$z$ PDF storage: Sparse representation

Use Gaussian and Voigt profiles as bases, need $N^2_{\rm original}$ bases

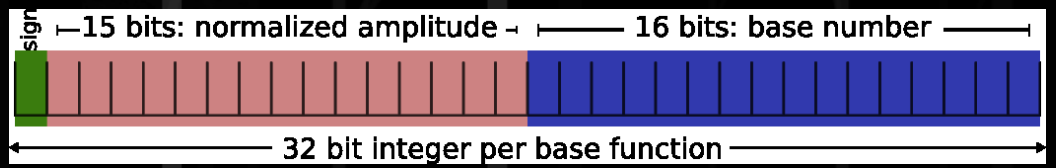
With only 10-20 bases achieve 99.9 % accuracy

Use 32-bits integer per basis, compression

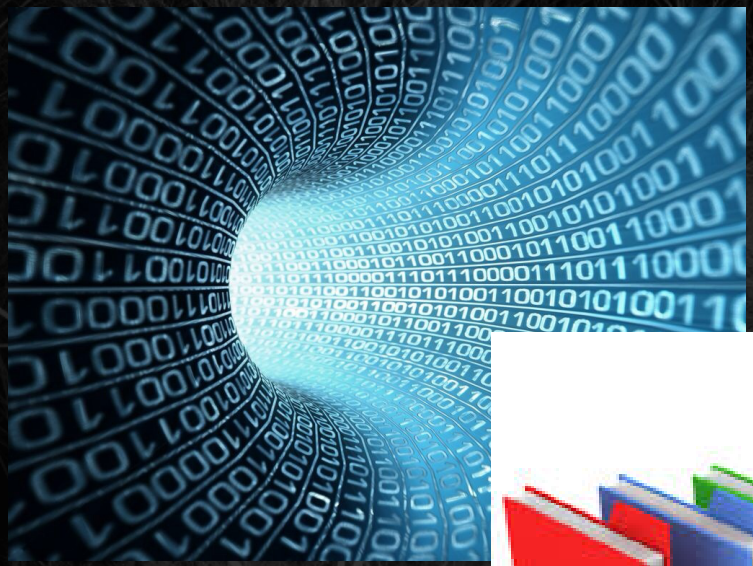
Store Multiple PDFs

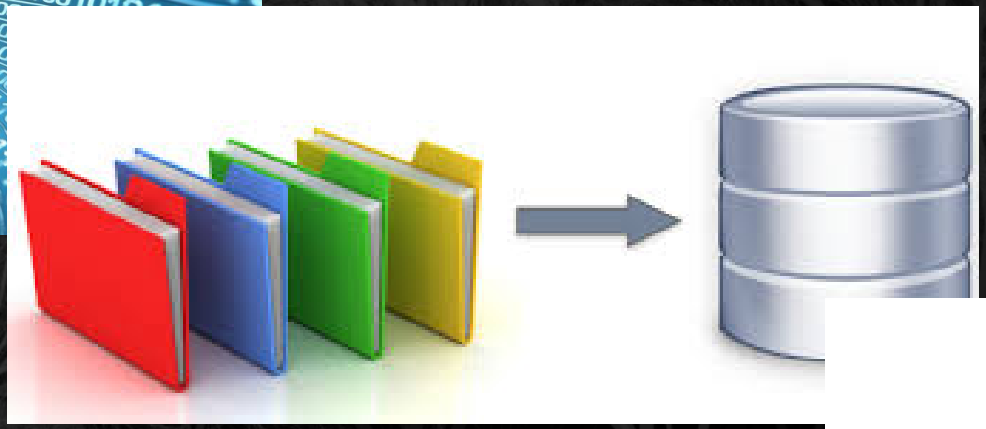



Carrasco Kind & Brunner 2014b (MNRAS, 441, 3550)



# How we can do it?





# Photo-z's in DESDM DB

Photo-z for ALL objects!

Y1A1 $\sim$ 130 M objects

Y2A1 expected to have more than 100 M objects

PDF methods requiere lot of space, sparse rep as alternative.

# New Photo-z package in Oracle

https://opensource.ncsa.illinois.edu/confluence/display/DESDM/Access+to+photo-z+from+DB

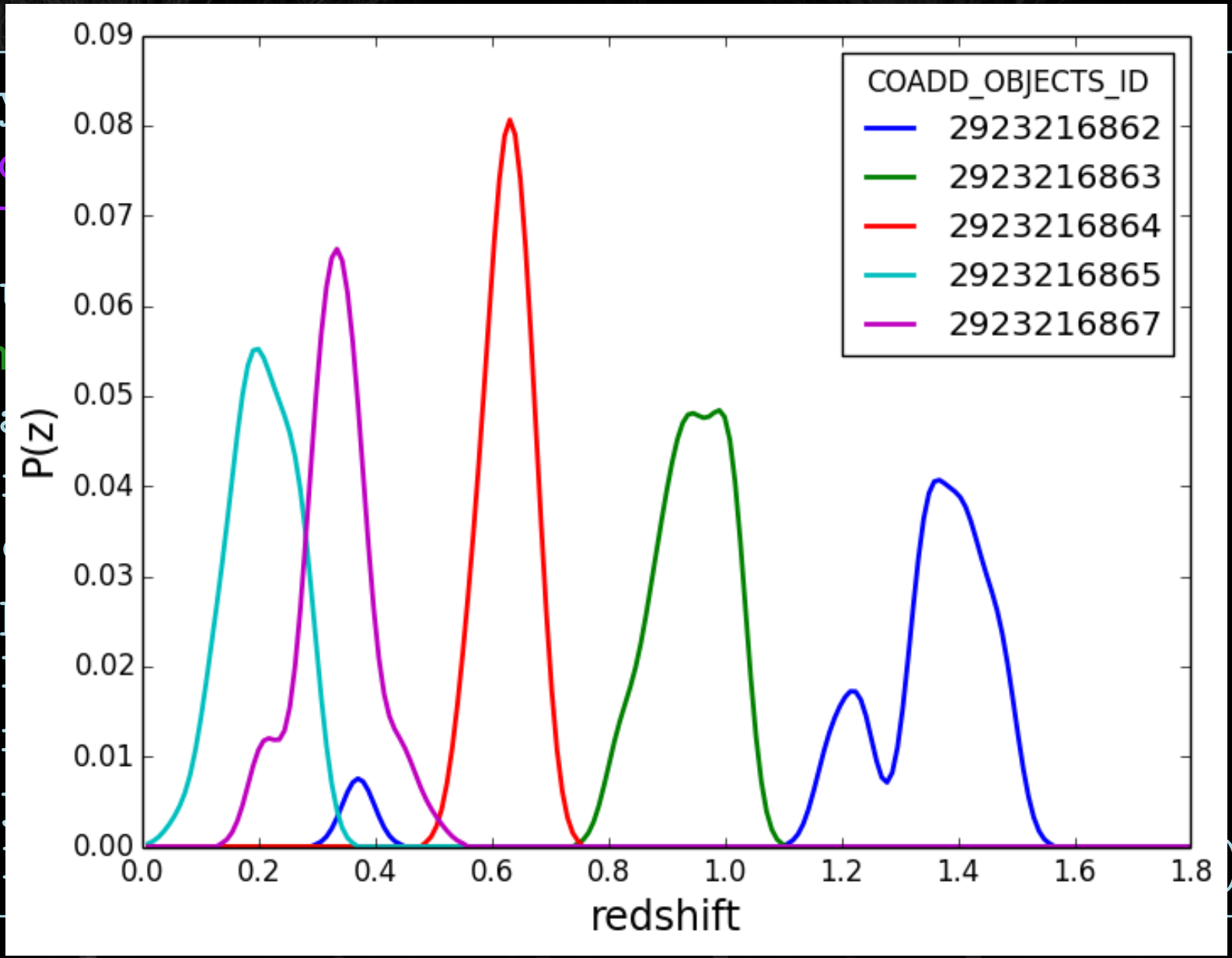
- New data types
  - * PFULL $\rightarrow$ 200-vector
  - * PSPARSE $\rightarrow$ 20-vector
  - * PFULL_TB (ancillary)
- New functions
  - * GET_PDF (PSPARSE TYPE)
  - * MAX (PFULL TYPE)
  - * MEAN (PFULL TYPE)
  - * PEAK (PFULL TYPE)
  - * MEDIAN (PFULL TYPE)
  - * SUM (PFULL TYPE)
  - * NZ aggregate function (select NZ() from ...)

# Getting some PDFs from DB

```
query="""
select COADD_OBJECTS_ID,TPZ from
PHOTOZ_PDF_SVA1_GOLD where rownum < 6"""
cc=cursor.execute(query)
#Handling and plot
df=ea.to_pandas(cc)
for i in xrange(5):
    cid=df.COADD_OBJECTS_ID.values[i]
    plt.plot(zbins,df.TPZ.values[i],
    lw=2,label=cid)
plt.xlabel('redshift',fontsize=17)
plt.ylabel('P(z)',fontsize=17)
plt.legend(loc=0, title='COADD_OBJECTS_ID')
```

# Getting some PDFs from DB

```
query
sele
PHOT
cc=c
#Han
df=e
for
plt.
plt.
plt.
```



# Now using Sparse rep.

```
query="""
select COADD_OBJECTS_ID,PHZ.GET_PDF(TPZ) as
TPZ from PHOTOZ_SPARSE_SVA1_GOLD
where rownum < 6"""
cc=cursor.execute(query)
#Handling and plot
df=ea.to_pandas(cc)
for i in xrange(5):
    cid=df.COADD_OBJECTS_ID.values[i]
    plt.plot(zbins,df.TPZ.values[i],
    lw=2,label=cid)
plt.xlabel('redshift',fontsize=17)
plt.ylabel('P(z)',fontsize=17)
plt.legend(loc=0, title='COADD_OBJECTS_ID')
```

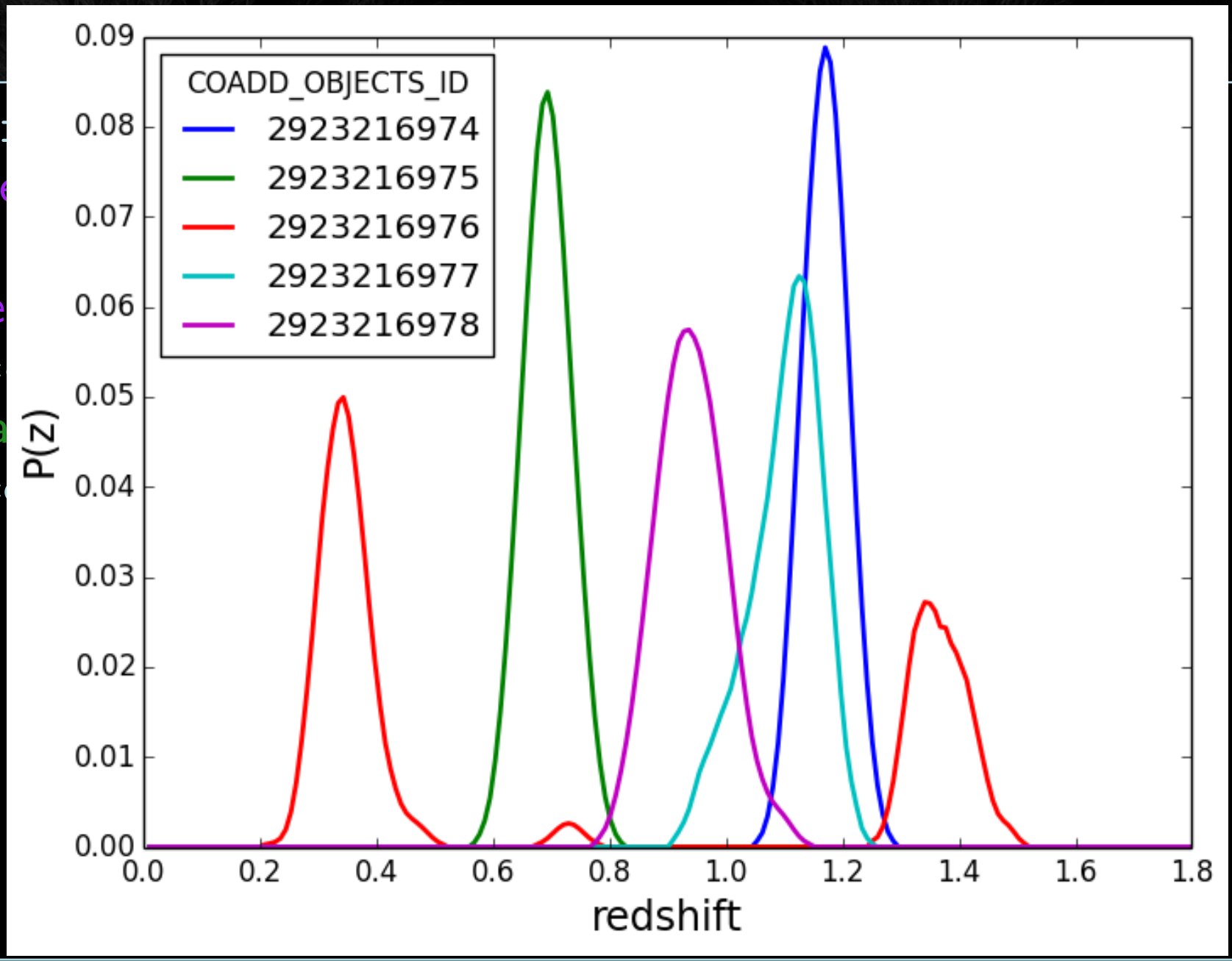
# Now using Sparse rep.

```
query="""
select COADD_OBJECTS_ID,PHZ.GET_PDF(TPZ) as
TPZ from PHOTOZ_SPARSE_SVA1_GOLD
where rownum < 6"""
cc=cursor.execute(query)
#Handling and plot
df=ea.to_pandas(cc)
for i in xrange(5):
    cid=df.COADD_OBJECTS_ID.values[i]
    plt.plot(zbins,df.TPZ.values[i],
    lw=2,label=cid)
plt.xlabel('redshift',fontsize=17)
plt.ylabel('P(z)',fontsize=17)
plt.legend(loc=0, title='COADD_OBJECTS_ID')
```

# Now using Sparse rep.

```
quer
sele
TPZ
whe
cc=
#Ha
df=
for

plt
plt
plt
```

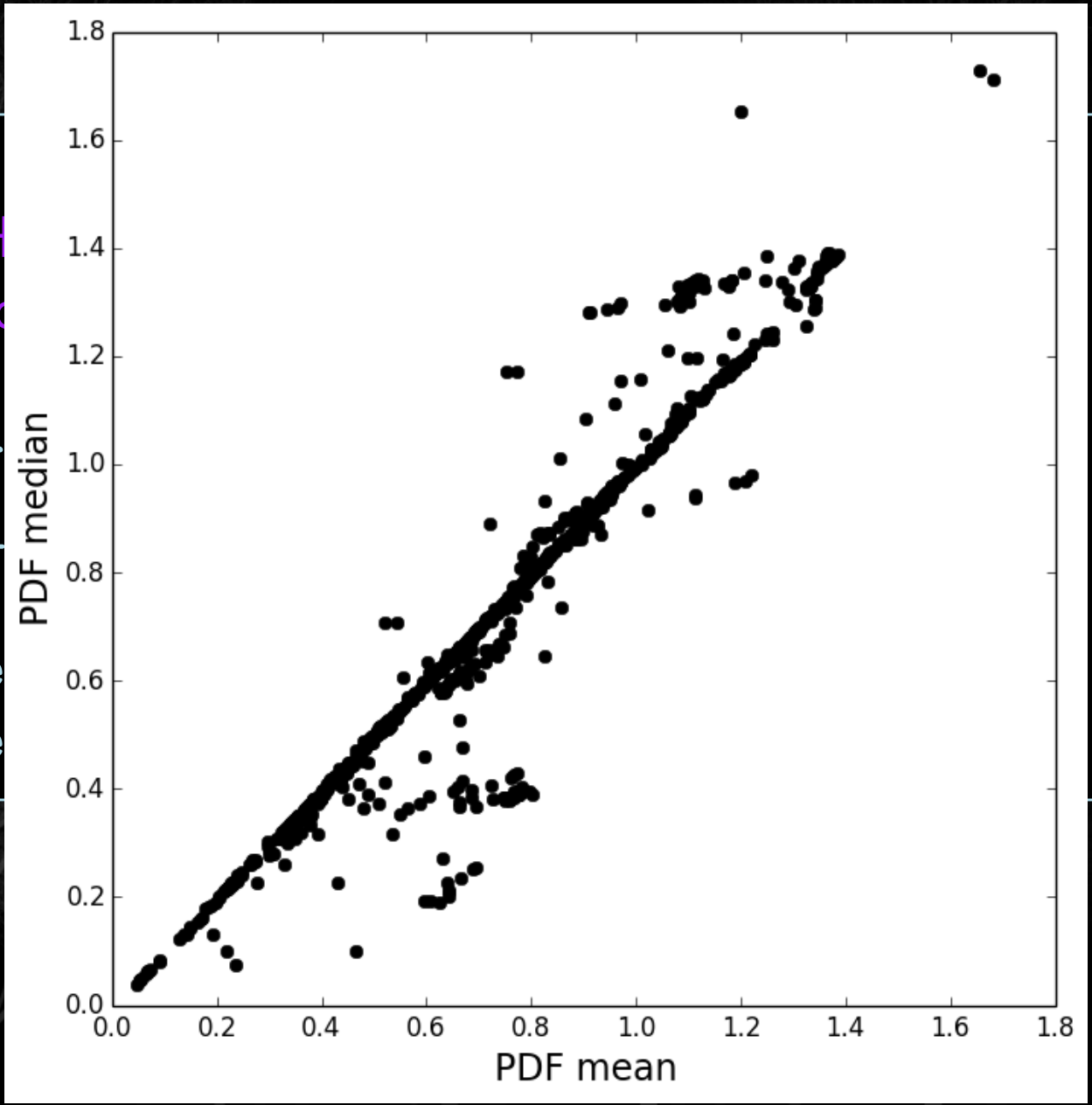


# Getting metrics on the fly

```
query="""
Select PHZ.MEAN(tpz) mean, PHZ.MEDIAN(tpz)
median from PHOTOZ_PDF_SVA1_GOLD
where rownum < 1000"""
cc=cursor.execute(query)
df=ea.to_pandas(cc)
plt.plot(df.MEAN,df.MEDIAN,'ko')
plt.xlabel('PDF mean',fontsize=17)
plt.ylabel('PDF median',fontsize=17)
```

# Getting metrics on the fly

```
query="""
Select PH                                    pz)
median fr
where row
cc=cursor
df=ea.to_
plt.plot(
plt.xlabe
plt.ylabe
```



# Stacking PDFs in DB cluster!

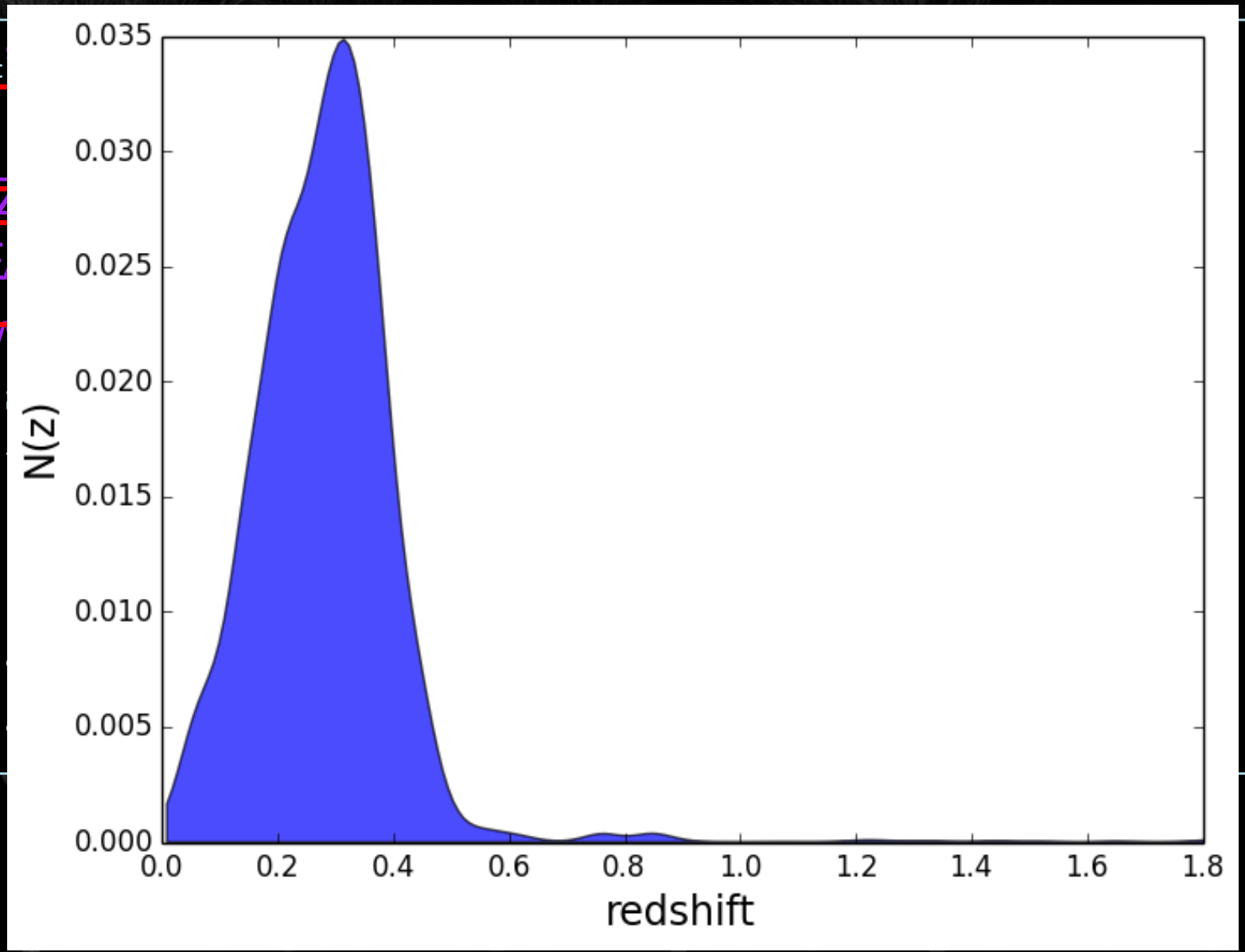
```
query="""
Select NZ(PHZ.TOTABLE(tpz)) as NZ from
PHOTOZ_PDF_SVA1_GOLD where
PHZ.MEAN(tpz) BETWEEN 0.1 and 0.4
and rownum < 100000"""
cc=cursor.execute(query)
df=ea.to_pandas(cc)
plt.fill_between(zbins,df.NZ.values[0],
     facecolor='blue',alpha=0.7)
plt.xlabel('redshift',fontsize=17)
plt.ylabel('N(z)',fontsize=17)
```

# Stacking PDFs in DB cluster!

```
query="""
Select NZ(PHZ.TOTABLE(tpz)) as NZ from
PHOTOZ_PDF_SVA1_GOLD where
PHZ.MEAN(tpz) BETWEEN 0.1 and 0.4
and rownum < 100000"""
cc=cursor.execute(query)
df=ea.to_pandas(cc)
plt.fill_between(zbins,df.NZ.values[0],
      facecolor='blue',alpha=0.7)
plt.xlabel('redshift',fontsize=17)
plt.ylabel('N(z)',fontsize=17)
```

# Stacking PDFs in DB cluster!

```
query=
Select
PHOTOZ
PHZ.ME
and rov
cc=cur
df=ea.
plt.fi
      f
plt.xl
plt.yl
```



# Conclusions

- Photo-z tables in DB!
- Access to photo-z is easier and coordinated
- Use sparse representation for PDFs
- Bring analysis (software) to DB!
- github.com/mgckind/SparsePz

# THANKS!

# Questions?

Matias Carrasco Kind
NCSA/UIUC
mcarras2@ncsa.illinois.edu
https://github.com/mgckind

# EXTRA SLIDES