Mining Massive Data Sets With CANFAR and Skytree

Nicholas M. Ball
Canadian Astronomy Data Centre
National Research Council
Victoria, BC, Canada

Collaborators

- David Schade (CADC)
- Alex Gray (Skytree and Georgia Tech)
- Martin Hack (Skytree)

... and many others

Me

Data miner who does astronomy

Astronomer who does data mining

Outline

- CANFAR
- Skytree
- Combining them: CANFAR+Skytree
- Using it
- Example Applications
- Conclusions

CANFAR

Canadian Advanced Network for Astronomical Research

- CADC's cloud computing system to provide a **generic** infrastructure for storage and processing
- Processing: 500 cores, nodes up to 6 processors and 32G memory (soon 256G)
- Storage: **VOSpace**, several hundred terabytes available, **mounted filesystem**
- User sees a virtual machine, on which one can install and run any Linux software, e.g., **almost all astronomy code**
- Run code interactively or in batch, via Condor

SKYTREE®
BIG DATA ANALYTICS

- 7 well-known data mining algorithms (next slide)

- **Fast** implementations: N^2 -> N

- Robust, proven accuracy (FASTlab) -> **publication-quality results**

- Academic and astronomy background

- Works as command line as part of one's analysis

- E.g., input ASCII data, output and visualize results

SKYTREE®
BIG DATA ANALYTICS

Algorithm	Description	Runtime Notes
`allkn`	All nearest neighbors	O(N) (naive: N^2)
`kde`	Kernel density estimation	O(N) (naive: N^2)
`svm`	Support vector machine	Data-dependent
`lr`	Linear regression	Data-dependent
`svd`	Singular value decomposition	Data-dependent
`kmeans`	K-means clustering	Data-dependent
`two_pt`	2-point correlation function	O(N) (naive: N^2)

- Powerful system: Skytree on up to 500 cores in parallel

- Install on any VM with **own code**

- Access to **VOSpace** storage as **mounted filesystem**

- Analogy to CANFAR itself: **generic infrastructure** to facilitate science

- Extends CANFAR's capability to enable data mining

How to Use CANFAR+Skytree

- Request a CANFAR account
- Register for a CADC account
- ssh to CANFAR, start a virtual machine
- Install Skytree on the virtual machine
- Add license server to path
- Run Skytree (next slide)

```
> Email canfarhelp@nrc.ca
> Website login + password
> ssh canfar.dao.nrc.ca
> vmcreate <myvm>
> vmssh <myvm>
> tar -zxf <tarball> from http://
www.skytreecorp.com
> export SKYTREE_LICENSE_
PATH=@login-server-ip-address:/
home/username/
.skytree/skytree-client.lic
```

Running Skytree

- Typical Skytree call looks like, e.g.:

```
> ./skytree-server allkn \
    --references_in=datasets/sdss100kx4.skytree \
    --k_neighbors=1 \
    --distances_out=distances.out \
    --indices_out=indices.out
```

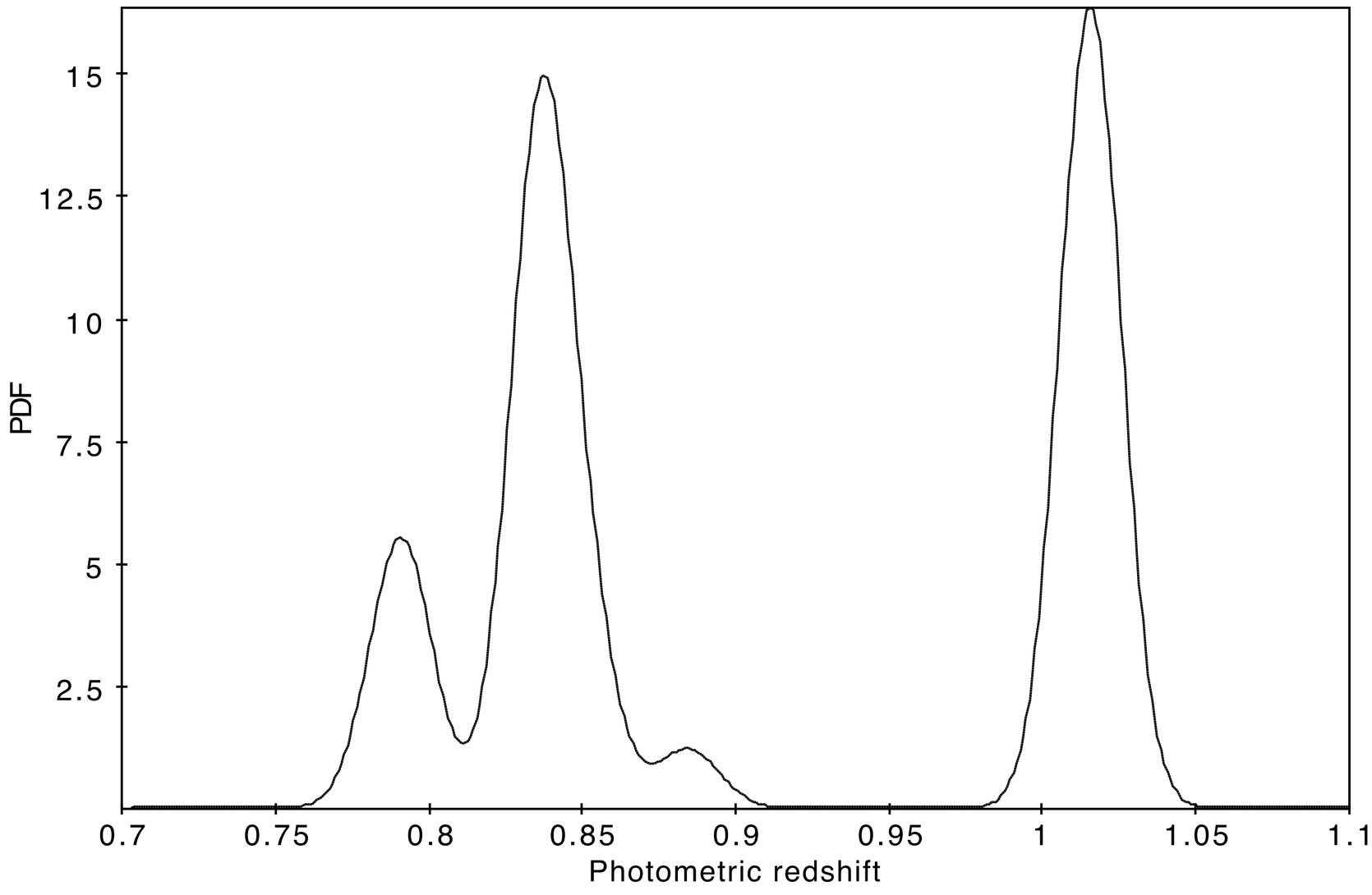
- Interactive, or batch (up to 500 cores) as part of your data processing

Consultation

- The aim of the system is to enable better science
- If you have a problem to solve, send us an email, and we'll work with you
- My background is astronomy, but I also know data mining

Example: Photo-zs for CFHTLS

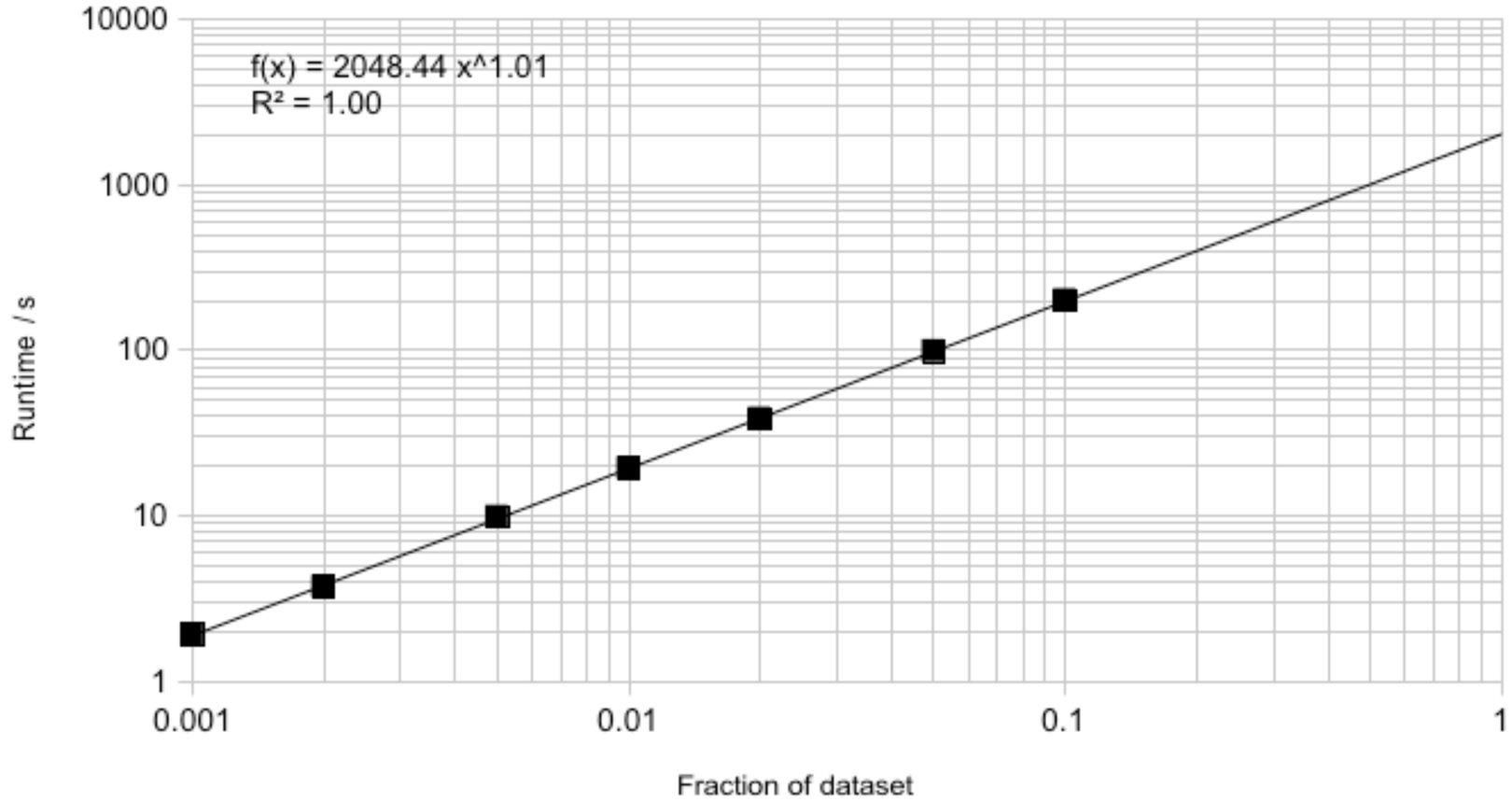
- Own science interest is the galaxy luminosity function
- But legacy value of photo-zs
- Skytree `allkn` allows generation of full PDFs via perturbing inputs (Ball et al. 2008); and CANFAR allows comparison to template-based (Le Phare)
- Done for ~130 million CFHTLS galaxies (~26m x 5 detection passbands)
- 100 perturbations: create and handle a catalogue of ~13 billion objects
- -> **CANFAR+Skytree can process LSST-sized datasets**



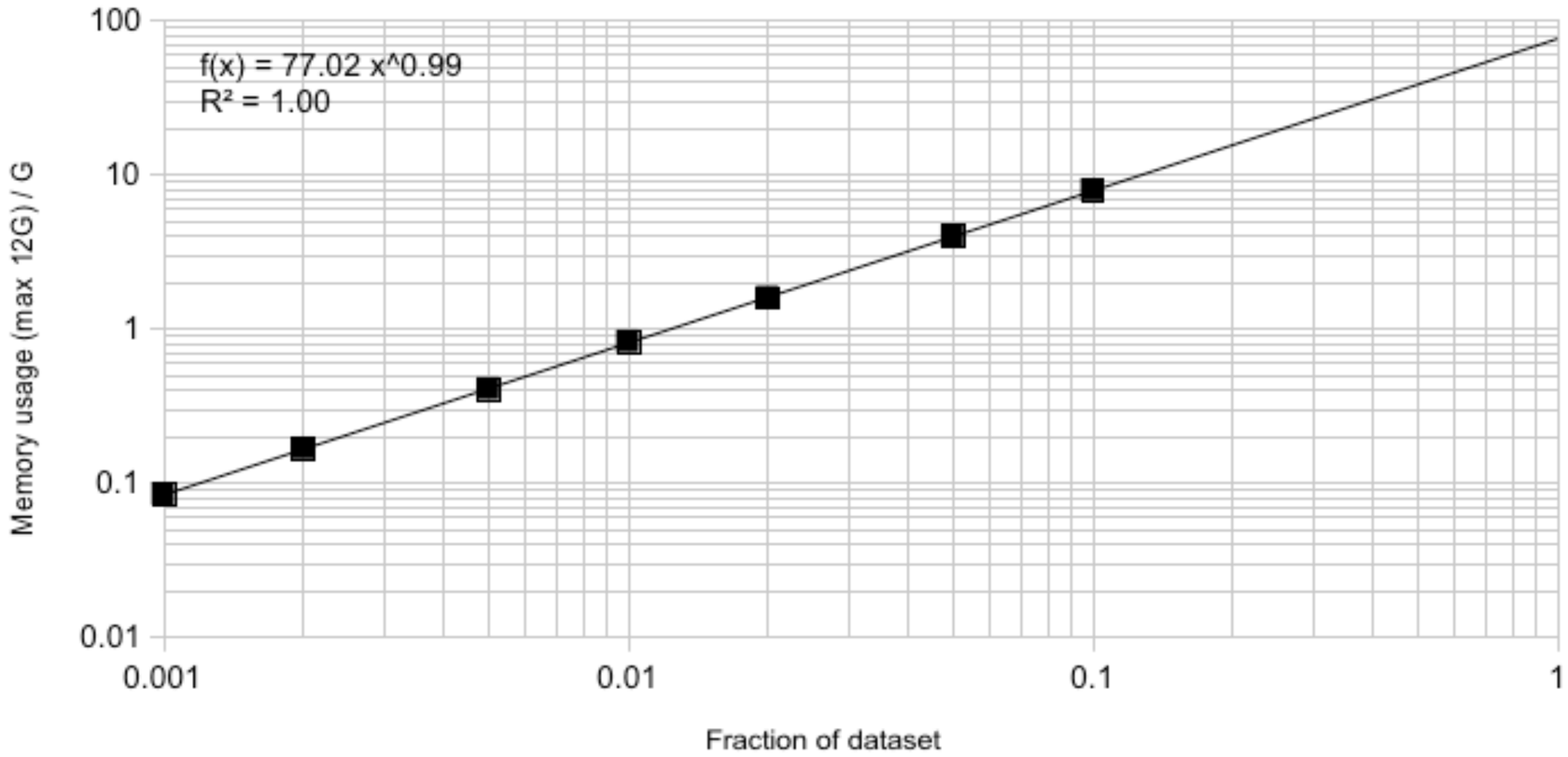
allkn photo-z instances fitted with **kde**

Example: Skytree Scaling

- Show that Skytree scales as claimed on real astronomy data
- Do we really get N^2 -> N?
- Compare to open-source alternative, R
- Run algorithms on large catalogues: 100 million objects or greater
- E.g., 2MASS, CFHTLS, WISE, etc.
- Do useful investigations, e.g., find outliers



Work in progress



Work in progress

Conclusions

- CANFAR: storage, processing, analysis, **generic**, with **own code**
- Skytree: fast, robust -> **publication-quality results**
- CANFAR+Skytree: Skytree up to 500 cores, combine with own code, access to **VOSpace**
- To get started, email canfarhelp@nrc.ca (or talk to me!)
- For more information: Poster, or https://sites.google.com/site/nickballastronomer

We encourage interested users!