CANFAR

Canadian Advanced Network for Astronomical Research

KDD-IG and CANFAR

Nick Ball

CADC, Herzberg Institute of Astrophysics, Victoria

Nick.Ball@nrc-cnrc.gc.ca

https://www.astrosci.ca/users/NickBall

University of Victoria

University of British Columbia

CADC/CCDA

NRC·CNRC From Discovery to Innovation... De la découverte à l'innovation...

canarie

CANFAR

Canadian Advanced Network for Astronomical Research

KDD-IG wants to deploy practical data mining algorithms:

“We will develop and test scalable data mining algorithms and the accompanying new standards for VObs interfaces and protocols, so that these algorithms can be discovered and used transparently within VO science workflows or in standalone data exploration applications.”

(KDD-IG draft charter)

CANFAR

Canadian Advanced Network for Astronomical Research

CANFAR provides an infrastructure:

“The Canadian Advanced Network for Astronomical Research (CANFAR) is a project ... to provide the delivery, processing, storage, analysis, and distribution of astronomical datasets of unprecedented size. ... The project builds on CADC's existing infrastructure to provide IVOA-compliant tools and services for astronomers, and access to Cloud Computing on the Compute Canada Grid, via a Virtual Machine environment.”

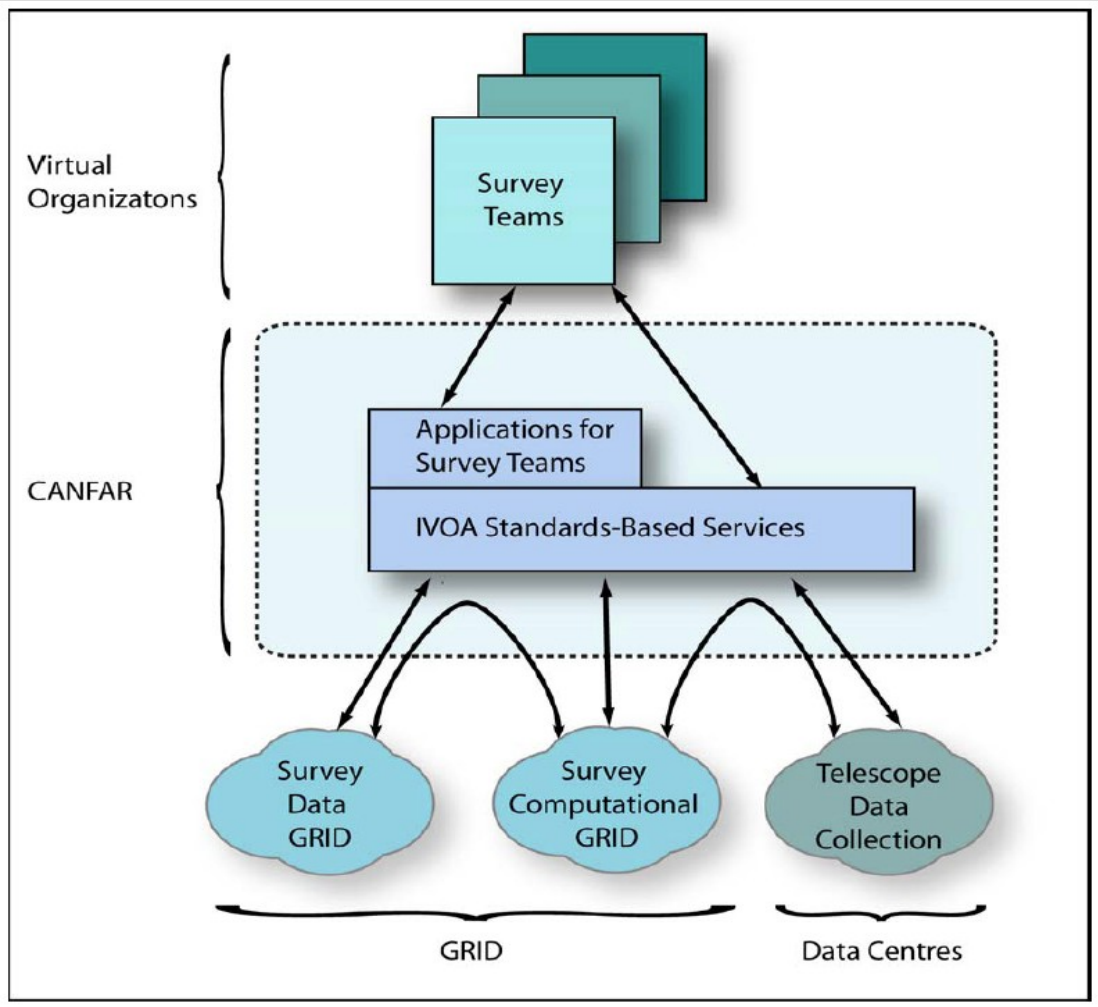
(CANFAR Statement of Work 2008)

CANFAR

Canadian Advanced Network for Astronomical Research

CANFAR infrastructure

http://www.astro.uvic.ca/~canfar



CANFAR

Canadian Advanced Network for Astronomical Research

But data mining algorithms have been N^2

Need NlogN to make them tractable on large datasets

Are now libraries available that do this

CANFAR

Canadian Advanced Network for Astronomical Research

E.g. Gray @ IJCAI '09

Computational complexity using fast algorithms

- **Querying:** nearest-neighbor **O(logN),** spherical range-search **O(logN),** orthogonal range-search **O(logN),** contingency table
- **Density estimation:** kernel density estimation **O(N) or O(1),** mixture of Gaussians **O(logN)**
- **Regression:** linear regression **O(D) or O(1),** kernel regression **O(N) or O(1),** Gaussian process regression **O(N) or O(1)**
- **Classification:** nearest-neighbor classifier **O(N)**, nonparametric Bayes classifier **O(N)**, support vector machine **O(N)**
- **Dimension reduction:** principal component analysis **O(D) or O(1)**, non-negative matrix factorization, kernel PCA **O(N) or O(1)**, maximum variance unfolding **O(N)**
- **Outlier detection:** by robust L_2 estimation, by density estimation, by dimension reduction
- **Clustering:** k-means **O(logN)**, hierarchical clustering **O(NlogN),** by dimension reduction
- **Time series analysis:** Kalman filter **O(D) or O(1),** hidden Markov model, trajectory tracking
- **2-sample testing:** n-point correlation **O(N^{logn})**
- **Cross-match:** bipartite matching **O(N) or O(1)**

http://www.astro.caltech.edu/~george/AIworkshop/presentations.html

CANFAR

Canadian Advanced Network for Astronomical Research

-> install NlogN - capable libraries on the CANFAR infrastructure

Develop using CADC expertise (includes Astronomy+CS)

If the libraries need developing, their authors, KDD-IG activities, and charter can help guide this

CANFAR

Canadian Advanced Network for Astronomical Research

Example Science:

What is the faint end slope of the galaxy luminosity function in the Virgo Cluster?

Use the Next Generation Virgo Survey (NGVS, a CANFAR project)

The LF faint end constrains the baryonic component of the concordance ΛCDM cosmological model

NGVS expected survey size: 2.6T

CANFAR

Canadian Advanced Network for Astronomical Research

Extract catalogue of objects: SExtractor, template matching, KSOM, MARSIAA, etc.

Classify them as star/galaxy/other: PSF, SVM, etc.

Are they in Virgo or the background: EM algorithm?

Fit functions to the LF: SWML, etc.

Photometric redshifts: kNN, full PDF

N-point correlation function

Detailed profile fitting for morphology: GALFIT

CANFAR

Canadian Advanced Network for Astronomical Research

The science is specific

But the tools are not

Could install as part of a basis set of software on CANFAR VMs

CANFAR

Canadian Advanced Network for Astronomical Research

Summary

NlogN data mining algorithms

+

CANFAR

=

New science!

University of Victoria

University of British Columbia

canarie