1. Provide a brief abstract for the project.

The Next Generation Virgo Cluster Survey (NGVS) is an approved Large Programme for the Canada French Hawaii Telescope (CFHT). The NGVS will use 771 hours of CFHT time (approx. 140 nights), spread equally over the 2009A-2012A semesters, to image the Virgo Cluster — the dominant mass concentration in the local universe and the largest collection of galaxies within ≈35 Mpc — from its core to virial radius, in five filters (*u,g,r,i,z*), to a point-source depth of $g' \sim 25.7$ mag (10σ) and a corresponding surface brightness of $\mu_{g'} \sim 27.7$ mag arcsec^{-2} (2σ). The NGVS represents the largest contiguous field ever observed at these depths.

The NGVS will supersede all previous optical studies of this uniquely important system, and, by leveraging the vast amount of data available at other wavelengths, will allow us to address a wide range of fundamental astrophysical questions, including: the faint-end shape of the luminosity function, the characterization of galaxy scaling relations over a factor 10^7 in mass, the cluster/intracluster medium/galaxy connection, and the fossil record of star formation and chemical enrichment in dense environments. Numerous ancillary projects will also be enabled, including a survey of the Galactic halo, the detection of Kuiper Belt Objects, the study of high-redshift clusters and galaxies, strong gravitational lensing, and cosmic shear measurements of the matter power spectrum on large scales

The NGVS will be the state-of-the-art optical survey of a low-redshift cluster environment for years to come, outperformed only by the 2025 completion of the LSST 10-year 20,000 deg^2 survey.

2.Describe why this project is appropriate to be included as one of the CANFAR science projects.

The following points make the NGVS suitable as a CANFAR project:

- Significant resources, both in terms of storage space (0.5 TB/night, both long and short term) and CPU, are required for each night observing (see next section)
- Archiving of all NGVS data (raw, Elixir-processed, MegaPipe and Terapix stacks) and data products (all catalogues) will be handled by the CADC (efforts in this directions are already underway).
- The data and data products need to be distributed to collaborators across Canada (U. de Montreal, UofT, Queens U., McMaster U., St.Mary's U., UBC, UVic, HIA), France (Paris, Lyon, Marseille, Strasburg), and locations in the UK, Italy, China, Chile and the US (including Hawaii).
- Because of the varied nature of the science goals, each of which requires optimized and dedicated data processing tools, the CANFAR infrastructure will significantly increase the productivity of the project, by means of efficiently allowing for selection of cutouts, catalogue queries, etc. for further dedicated processing.
- For the same reasons given above (we will be looking at objects spanning a huge range in sizes, mass, morphology, etc.), building an entirely automated pipeline is impossible. Even in the later stages of the survey, interactive monitoring of data processing will be necessary.
- The NGVS offers a wealth of synergistic opportunities with many existing, on-going and planned surveys of the Virgo Cluster at longer and shorter wavelengths, requiring cross-referencing of multiple databases.

We note that at the present time, the NGVS is 0% funded, a fact that is negatively impacting the PI's sleeping patterns. The PI is also ineligible to apply directly for any source of funding.

3.Describe the amount of telescope time allocated to the project, and the anticipated start and end dates for data collection. Describe any non obvious risk factors in the data collection which have a significant probability of occurring.

The NGVS was allocated 771 hours (~140 nights) of dark/gray time, the second largest time allocation in the 29-year history of the CFHT (after the CFHTLS). Data acquisition will start in 2009A and end in 2012A (spring semesters only). Approximatively 193 hours will be allocated each semester, taking an average of ~40% of each MegaPrime run between mid-January and mid-May. We note that data for a pilot project (covering 4 square degrees centred on M87) are already in hand, so CANFAR resources could be put to use effective immediately.

Barring a catastrophic instrument failure or exceptionally bad weather, there are no obvious risk factors that can affect the completion of the survey or the achievements of its science goals. The technical feasibility of the survey has already been demonstrated as part of a pilot project approved, in PI mode, for the 2008A semester. All pilot project images have been stacked and are being used to optimize the observing strategy of the full survey. Note that there is no specific time series built into the survey, i.e. there are no aspect of the science that would be impacted by changes in the scheduling.

4. Estimate the expected resources required for the project, in units of night of data collected.

As a baseline, each NGVS field will comprise 85 GB of raw data, 85 GB of Elixir-processed data, 34 GB of stacked images and 34 GB of weight maps (17 GB for each MegaPipe and Terapix processing). The data storage requirements for the stacked and weight images will easily double since we expect different science goals to require different stacks (for instance produced using different procedures for subtracting the sky or combining the images). For a single night observing, we expect an average of 114 GB of raw data, 114 GB of Elixir processed data, and 45.6 GB of Megapipe stacks (science data and weight images). Most science application will require additional image manipulation (median filtered / wavelet subtracted images, smoothed images, etc), for which we estimate a minimum of 200 GB of additional data. In total, therefore, we will require an average of 0.5 TB of disk space for each night observing, not including the Terapix stacks. [Effectively, most application will not need the raw data — which can be archived separately — reducing the amount of working space to 0.36 TB per each night observing]

At this stage, I don't think I can estimate the total amount of CPU hours required nor the minimum amount of memory. As an example, however, I am currently doing source extraction on full stacks which have been median-filter subtracted. For each stack, the entire process (median filtering, etc..) takes 1.1 hours on my Mac (2X32 GHz Quad-Cote Intel Xeon,16 GB 800 MHz DDR2 F8-DIMM). Running the same process on the data acquired each night (on average 1.3 fields, 5 bands each, short and long exposures), would require 14.8 hours. This is only a small part of the processing needed. Incidentally, on my Mac, IDL cannot allocate enough memory to handle full stacks.

In terms of timing, most of the resources will be required during the first two years of the survey (2009-2010). After that, it is hoped that the machinery will be well oiled and running smoothly. I would estimate that 80% of the resources will be needed during the first two years, and 20% during the last two (but of course this assumes that only "maintenance" will be required beyond 2010).

The NGVS already has a dedicated webpage, mailing list and blog, so it is not expected that there will be much need of CANFAR resources for this purpose. As detailed in the next section, however, we will need significant resources for database and catalogue queries.

5. Describe the major processing steps anticipated to be required.

At the end of this document, I will append a (perhaps exceedingly) detailed description of the storage/data manipulation tools that will be needed at the onset of the survey. These tools are indispensable for effective handling of the data. Furthermore, they have wide applicability (not only for MegaPrime data, but for any wide field imaging data, at CFHT and other telescopes), and will be easily adapted into a general, all-purpose tool that will be of service to the entire astronomical community.

The KBO science requires time critical processing to allow for follow-up of the objects and orbit determination. To this aim, processing should be completed within 3 days of acquisition. [JJ Kavelaars will have a more detailed discussion of this point in his writeup].

The main science goals of the NGVS do not strictly require time-critical processing. However, all data are immediately in the public domain, and prompt processing is necessary to ensure that the team can be competitive in achieving the science goals of the survey. Once the survey is up to speed, our goal is to complete all of the basic and advanced data processing (stacking, catalogue extraction, automated structural characterization of galaxies, etc..) in a semiautomated way for the data acquired during each MegaPrime run before the end of the next run (i.e. on a ~30-day timescale).

For the first semester of the survey, this process will not be automatically initiated and will require constant human intervention while the processing pipeline is being built and optimized. For the remaining three semesters of the survey, it is hoped that most processing can be initiated automatically and with minimal human intervention. It will still be necessary to manually check the Elixir-calibrated images as they become available, and supervise all advance processing (for instance, the stacking process, the source catalogues, etc.).

A. ACQUISITION AND BASIC PROCESSING

1. monitor progress in data acquisition.

There should be a dedicated webpage containing a chart of all fields expected to be observed. The chart should be updated daily to show fields that have been observed; for instance, fields to be observed could be gray and fields observed could be highlighted in a color corresponding to the filter used. Moving the cursor over the chart should display the name of the fields.

Within the same web interface, a second option should be given to use pull down menus to produce a list of all fields observed with given parameters (e.g. all fields observed in a given filter, all fields observed in photometric conditions, all field observed in a specific seeing range, or a combination of those).

2. Provide interactive access to basic information of raw and Elixir processed data.

The chart mentioned in 1. should be interactive: by clicking on an "observed" field, a list should be automatically produced and displayed giving details of the observations (date in which the images were taken, exposure times, average seeing as measured by Elixir, etc.) and including the rootnames of corresponding raw and Elixir-processed data. Additionally, clicking on a field should produce a thumbnail showing the MegaPipe stack.

3. Provide interactive access to the raw and Elixir data.

Checkboxes should be associated to the rootnames of the raw and Elixir processed data produced through steps 1 and 2. Selecting the checkboxes should automatically generate a list for retreival of the data from the CADC archive.

4. Provide interactive images for all observed fields.

Clicking on the MegaPipe thumbnail associated with each field (step 2) should open a full resolution images (e.g. jpeg) of the stack within a "Goggle map" zoomable and pannable interface. A box should be displaying the coordinate (RA and DEC) as the cursor is moved across the image.

5. Interactive display of data catalogue.

Upon displaying the full (jpeg) image of the field, the option should be given to identify within the image objects drawn from a series of catalogues (examples are the VCC, the SDSS, and the MegaPipe and TeraPix catalogue file). For each catalogue, qualifiers should be given so that the user can select objects with specific characteristics (for instance, only objects in a given magnitude range, or only galaxies with specific Hubble Types, etc..).

As an option, the list of objects displayed should be downloadable to a ASCII file.

6. Interactive access to full resolution fits images and cutouts

Within the interface discussed above (step 4.) an option should be given to download a full resolution fits image of the field displayed (i.e. either the full frame or a single cutout). A table should allow the used to select which type of files to download (raw data, Elixir processed data, MegaPipe stacks, Terapix stacks)

7. "on the fly" mosaicing of nearby fields.

By selecting multiple fields, these should be mosaiced together so cutouts straddling several fields can be selected.

B. ADVANCED PROCESSING

It would be desirable to provide an interface to run several different pieces of software for advanced processing, alone or in sequence (i.e. the output from one feeding the next). Existing software packages include:

1. Wavelet analysis (IDL/C++) -- background estimation, removal of large galaxies.

2. SExtractor (C) -- object identification, basic photometry and structural parameters

3. DAOPhot (F77) -- PSF fitting photometry for unresolved objects

4. Ellipse (IRAF) -- 2D isophotal fitting

5. GalFit (C) -- 2D galaxy decomposition, structural parameters

6. GIMP2D (IRAF) -- 2D galaxy decomposiion, structural parameters

7. Publicly available codes to measure photometric redshifts.

The first order analysis to be performed is the determination of photometric redshifts. This entails the following steps:

1. For each field, run SExtractor to provide a catalogue of objects. This will fail to identify objects against bright galaxy backgrounds;

2. Run the Wavelet analysis to provide an image "clean" of large galaxies, and allow detection of faint sources in areas with bright background;

3. Run SExtractor again on the "wavelet subtracted" images, and allow detection of objects projected again a bright backgound;

4. Mesh the two SExtractor catalogues;

5. Run the codes for determination of photometric redshifts, with the meshed catalogue as input.

Another typical scenario could be to build a catalogue of both background galaxies and globular clusters surrounding M87. Then I would need the following steps:

1. select a cutout around M87. Note that this will require mosaicing several MegaCam fields together;

2. Run SExtractor to provide a first separation between stars and galaxies;

3. Run the Wavelet analysis to provide an image "clean" of the M87 light and allow detection of faint sources close to the center of
the galaxy;

4. Run SExtractor again on the "wavelet subtracted" images;

5. Mesh the two SExtractor catalogues (the first used for objects in the outskirts, the second for the objects closer to the center);

6. Run DAOPhot to obtain more accurate photometry of the unresolved sources;

7. Run Ellipse, GalFit and GIMP2D on the resolved galaxies.

8. Generate output catalogues for all of the above, and store then in the project database, in querably form.

All of these steps should be controlled within a single interface, where parameters for each code can be selected and all codes run consecutively.