

Understanding 21st Century Astronomical Data Cubes

G. Jogesh Babu

Eric D. Feigelson

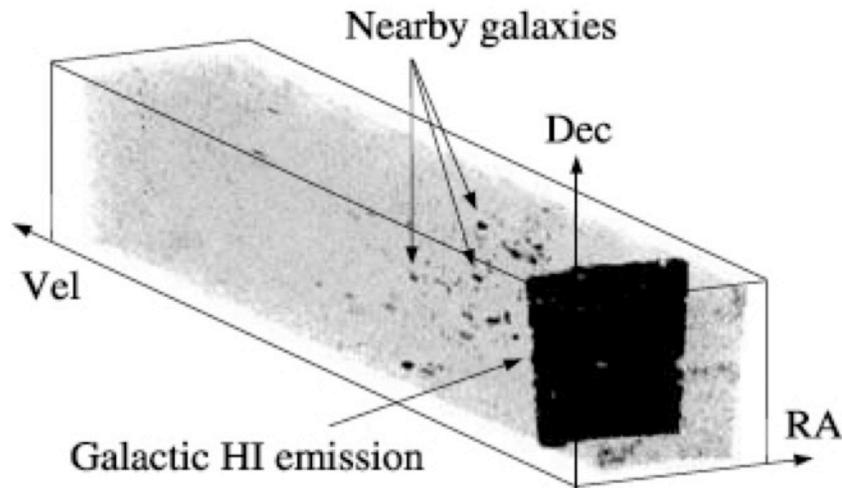
Center for Astrostatistics

Penn State University

Most 20-th century astronomical data have been 2-dimensional images or 1-dimensional spectra/time series. But 3-dimensional spectro-images and videos are becoming increasingly prevalent:

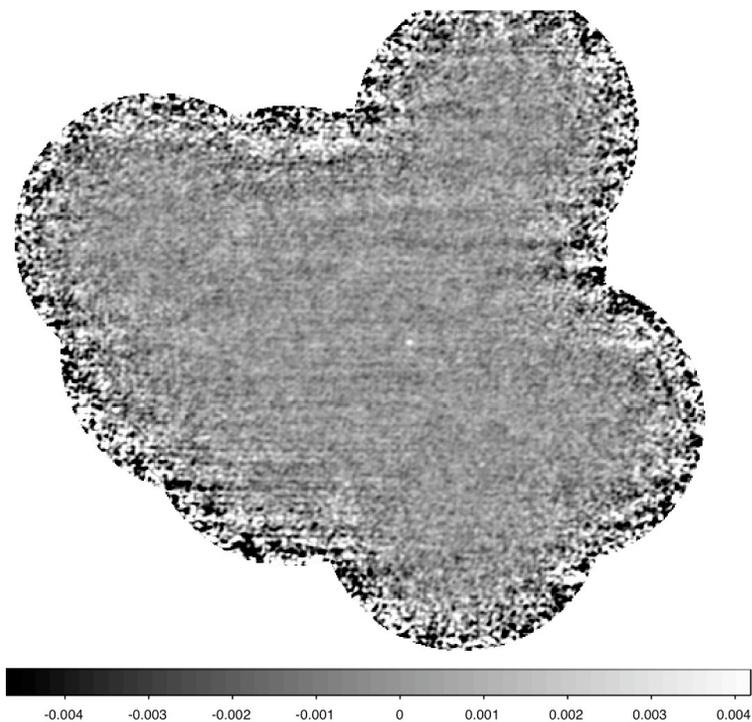
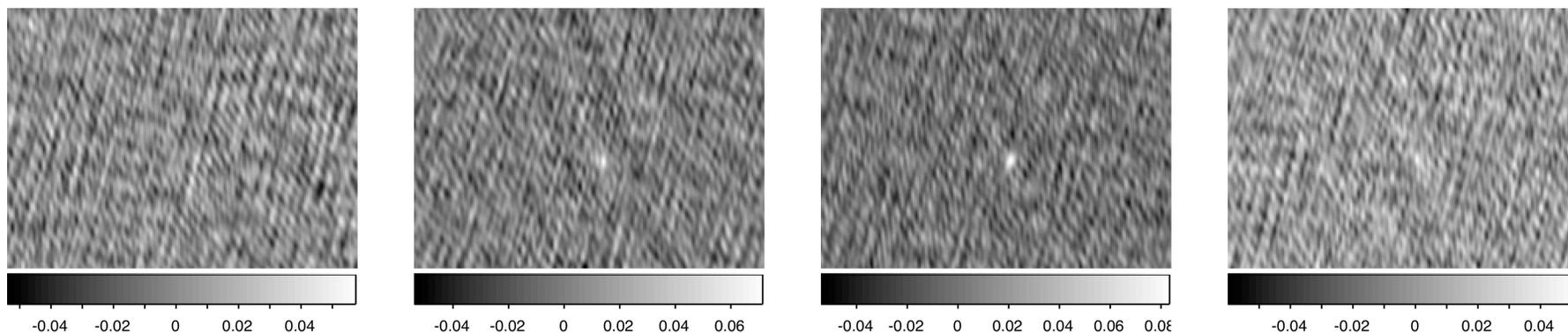
1. **Datacubes from radio interferometers**
(once restricted to 21-cm and molecular line maps, nearly all data from the EVLA and ALMA will be 3-dim spectro-image datacubes starting ~2013)
2. **Multiepoch visible-light surveys**
(Palomar QUEST, Pan-STARRS, LSST will produce huge datasets of time-image video-like cubes)
3. **Integral Field Units spectrographs**
(bundled-fiber-fed spectrographs give spectro-image cubes)

Bump hunting in radio astronomical data cubes



Three dimensional representation of a 21-cm HI data cube showing nearby galaxies (dark spots) and the Galactic Plane (dark sheet). Portion of the HI Parkes All-Sky Survey (HIPASS) catalog (Meyer et al. 2004)

The noise is often badly non-Gaussian and heteroscedastic (i.e. varies across the 3-dimensional image). Understanding the noise properties is particularly important for finding the faintest sources; this is called ‘bump hunting’.



Top: A subcube showing four adjacent channels of a large datacube of molecular maser emission in a Galactic star forming region, obtained at the Very Large Array (VLA). This shows the spatially correlated, non-Gaussian noise and faint sources common in radio datacubes. (Courtesy NRAO)

Left: A channel of a HI mosaic from the VLA illustrating heteroscedastic noise and ripples from mild RFI. (Courtesy J. van Gorkom)

The methodology widely used today for analysis of these radio datacubes is quite primitive:

- 2-dim sigma-clipping
 - too many false positives when noise is non-Gaussian,
 - too many false negatives for faint extended objects
 - insensitive to faint 3-dim sources
- 2-dim CLUMPFIND
 - robust sigma clipping with outlier rejection and edge correction
- 2-dim visual inspection
 - infeasible in era of terabyte-petabyte megadatasets
- TopHat (Australian Parkes surveys only)
 - median filter, multiresolution tophat filter weighted by local noise, features grouped in adjacent planes

The field has > \$1B of hardware investment and badly needs new data analysis methodology!

Datacubes are now emerging from integral field spectroscopy in optical and infrared astronomy

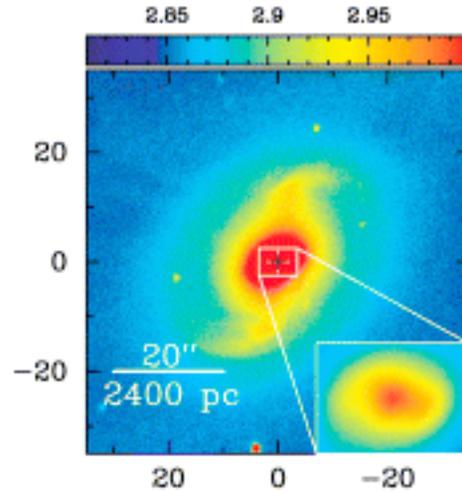
Integral Field Unit spectrographs are now at all major observatories:
GNIRS at Gemini-South, UIST at UKIRT, SparsePak & PMAS at WIYN,
VIMOS-IFU at VLT, GMOS at Gemini-North, IMACS at Magellan, TEIFU
at WHT, SAURON at WHT, OASIS at CFHT, GIRAFFE at VLT, VIRUS at HET

IFUs use lenslet arrays or optical fiber bundles to obtain spectra at $\sim 10^3$ contiguous locations. Goals include: galaxy kinematics, emission line galaxy maps, gravitational lensing, protostellar jet dynamics, protoplanetary disk gas excitation.

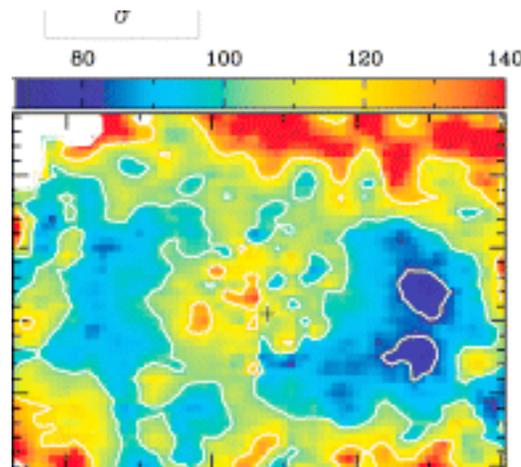
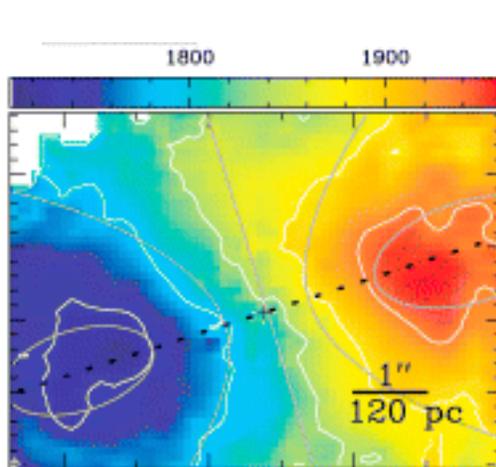
IFU data analysis is currently mostly based on low-dim projections of the datacube. Only a small fraction of the spatio-spectral information is extracted, and little effort is devoted to the low-S/N regime.

GMOS/Gemini study of spiral galaxy NGC 2273 kinematics

Barbosa et al. MNRAS 371, 170 (2006)



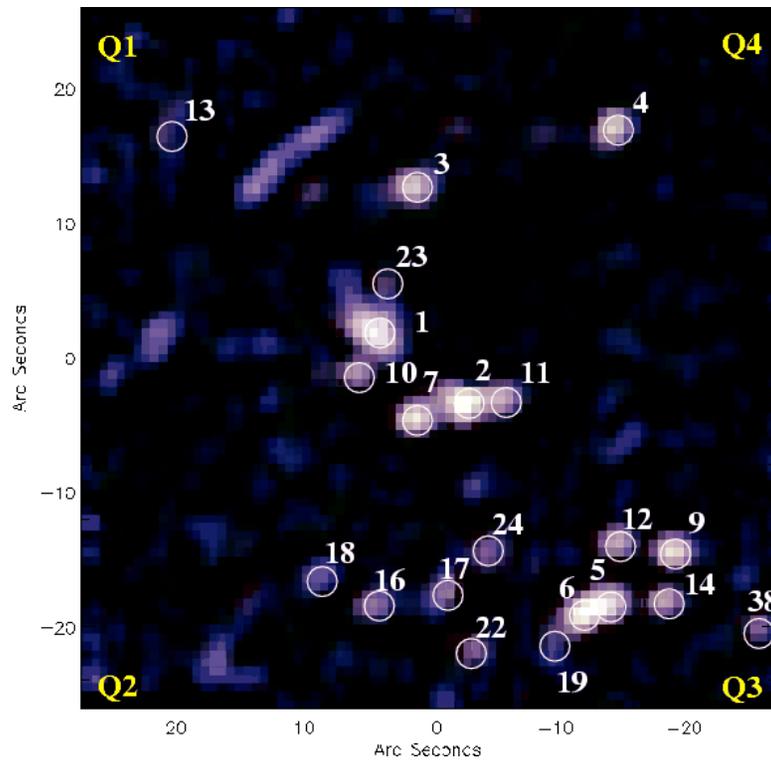
The panels here show 2-dimensional projections of the image intensity, Doppler shift and velocity dispersion. This reduces the spectral axis to two moments.



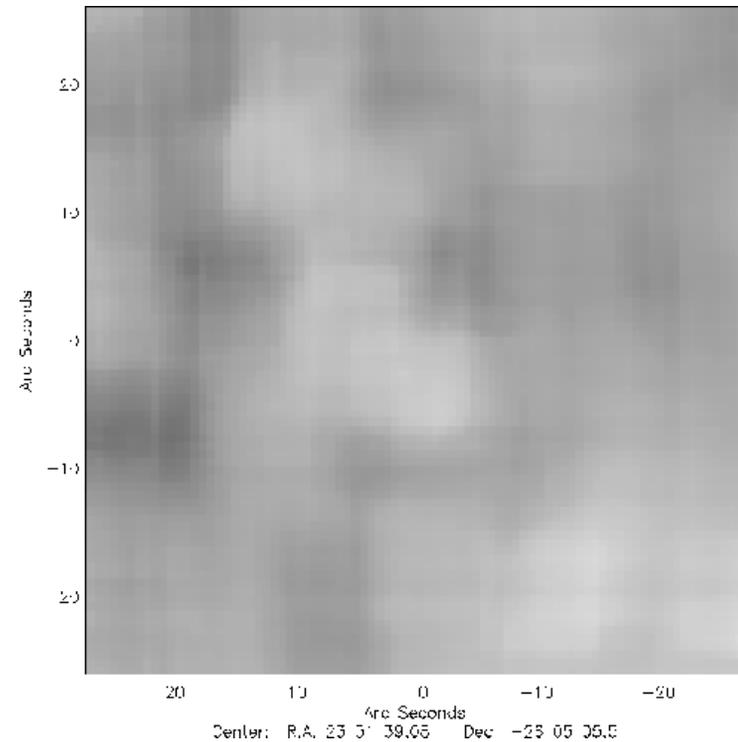
VLT VIMOS-IFU study of the gravitational lensing cluster Abell 2667

Covone et al. A&A 456, 409 (2006)

Color image. Here the spectral dimension is collapsed to 3 bands



Spatial variations in sky background noise in a 4 nm spectral window showing 10-fold heteroscedasticity



New procedures may involve:

- Robust rejection of bad pixels or planes
- Fully 3-dim signal detection using joint marginal quantiles in locally homoscedastic subcubes and/or non-parametric local regression
- Pixel-based identification of continuum/constant sources (rods in 3-dim) and line/transient sources (spots in 3-dim) using thresholds based on the local variance structure and autocorrelation between spectral/temporal planes
- False Discovery Rate control for false positives
- Image segmentation to unify adjacent hits
- Active contour of resulting 3-dim structures, including non-convex topologies

Local 3-dim subcube
for source detection

