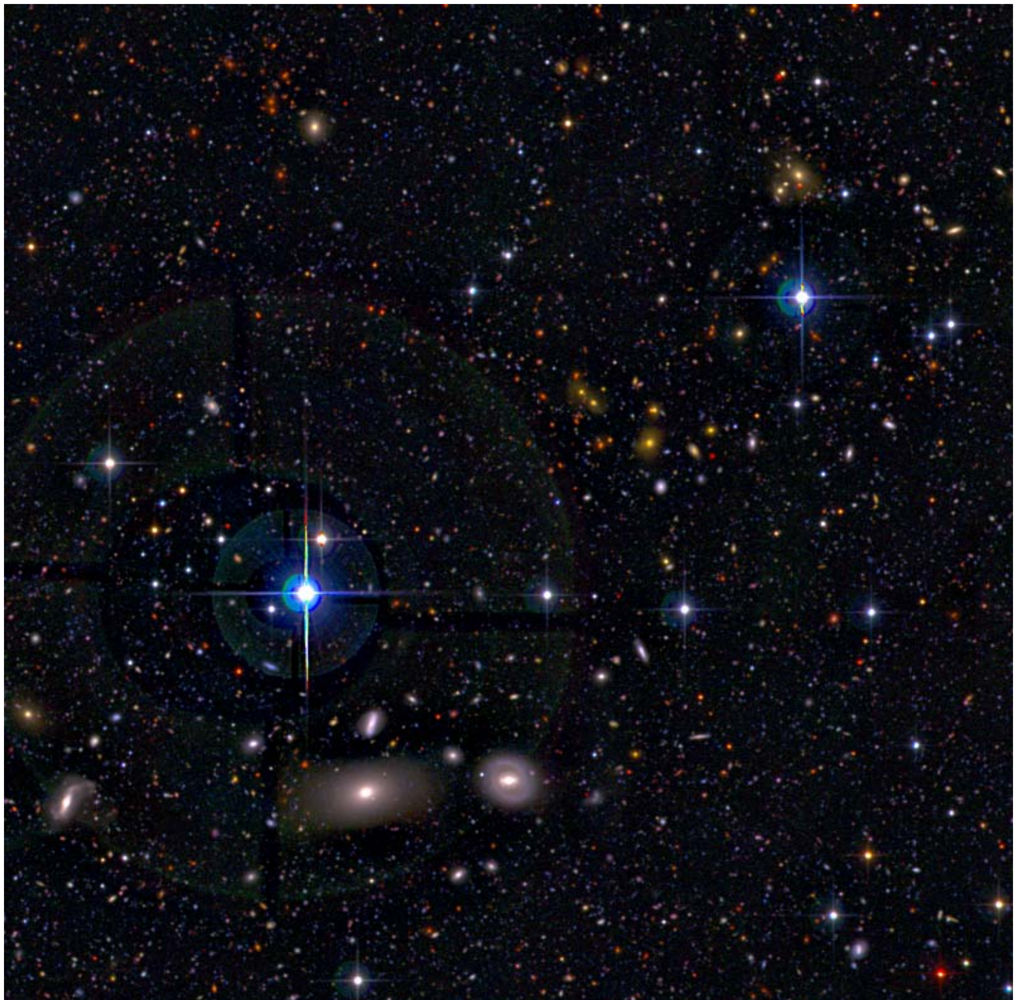
The making of visible/near-IR galaxy catalogs

E.Bertin, (IAP & Obs. de Paris/LERMA)

Outline



- Detection algorithms
- Multi-spectral analysis
- Measuring galaxy fluxes
- Measuring and classifying galaxy shapes

TERAPIX

What’s in an astronomical image?

S/N of sources between -6 and +100dB/pixel

Isophotal footprint of objects: from 1 to 10^6 pixels

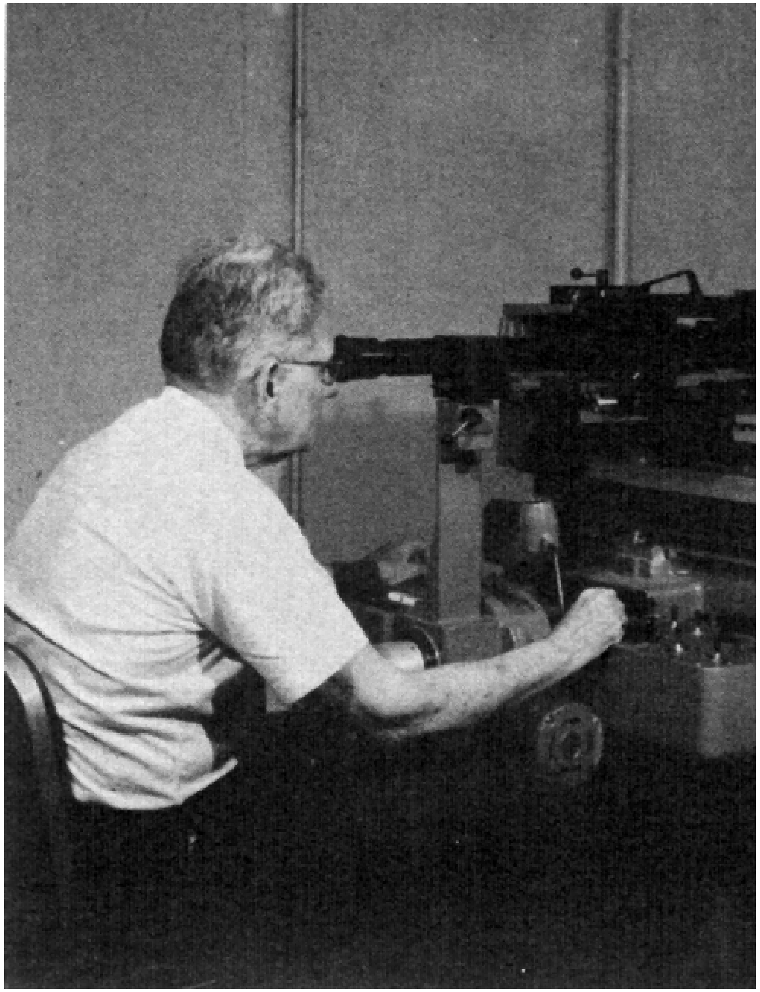
Image artifacts:
Halos
Detector blooming
Diffraction spikes
Cosmic-ray hits

“Almost stationary” Gaussian+Poisson noise correlated on small scales + large scale gradients

Most detectable sources are faint, barely resolved galaxies

Unsaturated stars can be used to map the variable PSF

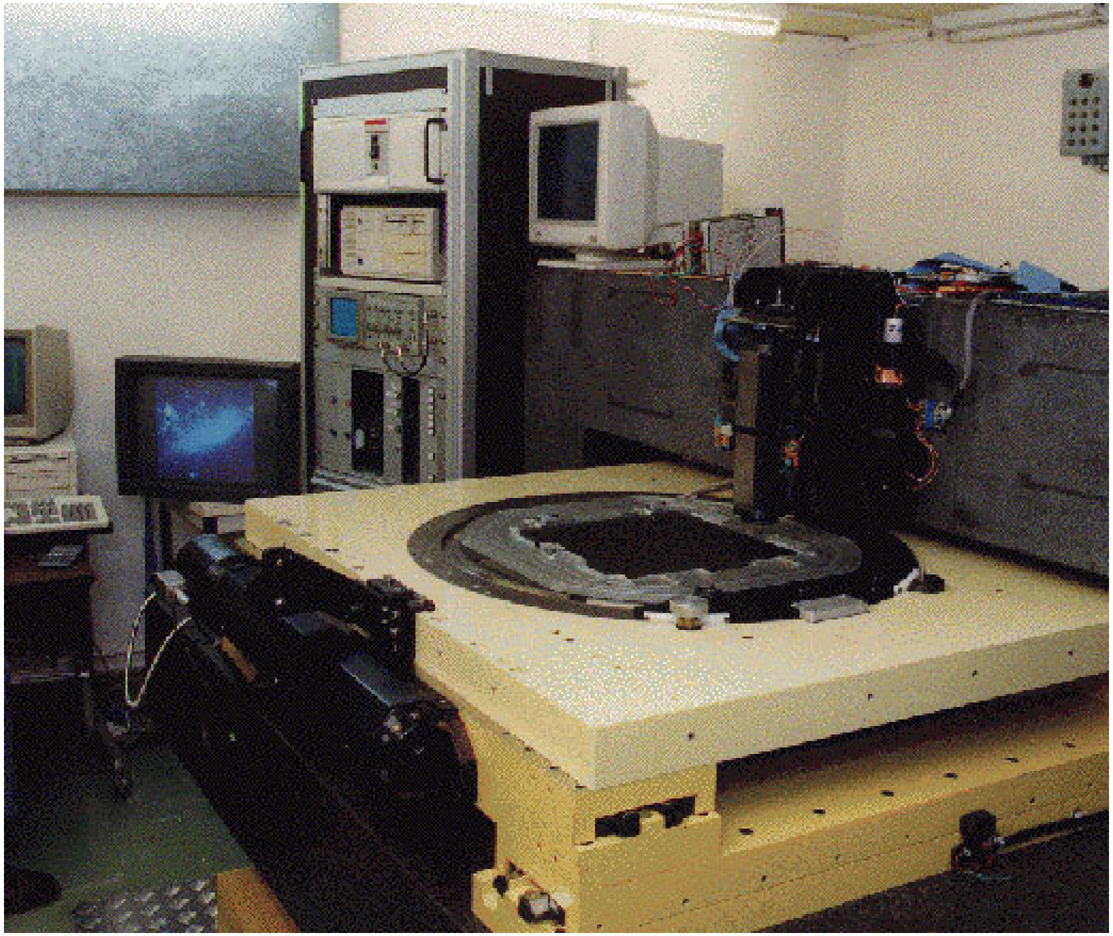
A bit of history…



- Until the end of the 60's, all astronomical sources where detected by eye on (photographic) images
  - Many astronomical catalogs where sources were detected and even measured "by eye" were still use 10 years ago (SAO, Zwicky,…).
    - ☹ Huge and tedious task. The "Schmidt problem" (Fellgett 1970): how to extract the tremendous amount of information stored on Schmidt plates?
    - ☹ The completeness and reliability of catalogs based on detections done by eye is variable and poorly defined.
    - ☹ Selection effects (especially for galaxies) ruin the benefit of large numbers: statistical studies are affected by large biases.

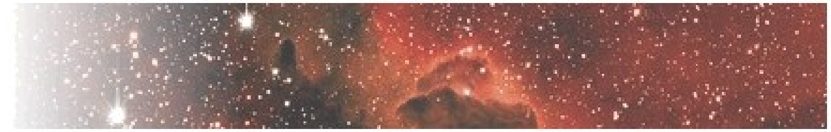
A bit of history (2)…

- End of the 60's / beginning of the 70's: the first automatic plate scanning machines are put into service (APS in Minneapolis, GALAXY in Edimburgh):
  - Source extraction is performed by dedicated hardware or supercomputers
  - Despite the high costs, the benefit of investing in automatic processing (in $ per source) for surveys is quickly realized.
- More machines are built around the world in the 70's and 80's: ESO S-3000, COSMOS, APM, MAMA, SUPERCOSMOS. Many of them are still in operation in 2006.



A bit of history (3)…

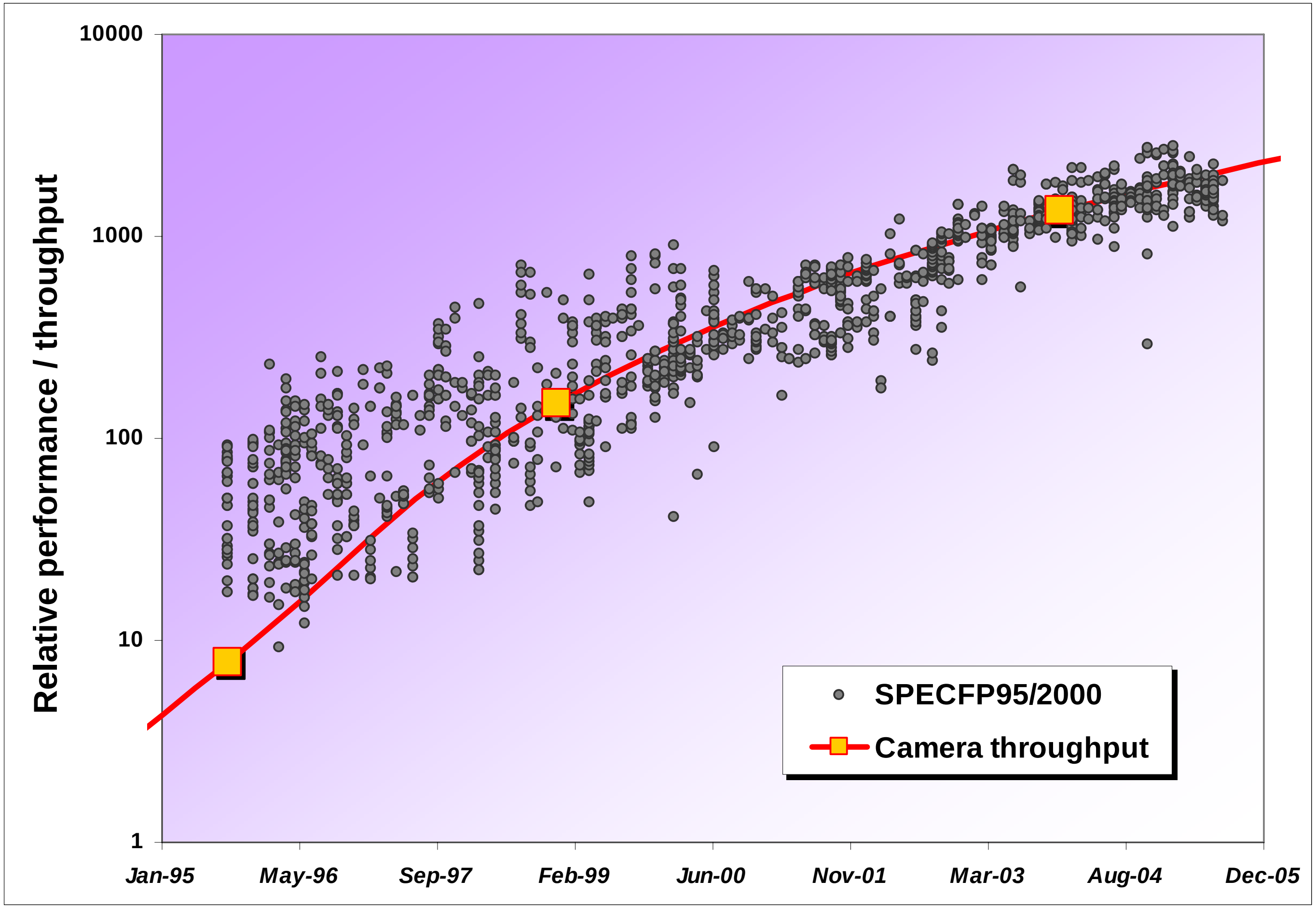
- 80’s and 90’s: small computers have sufficient resources now to run “stand-alone” source extraction software :
  - (crowded) star-field analysis software: DAOPhot, DoPhot, Romaphot, Inventory….
  - Galaxy extraction software: FOCAS, PISA, SExtractor,…
- Developing and testing software with a predictible, robust behaviour takes a long time
  - Until very recently most source extraction packages available were originally designed for photographic scans (low dynamic range, homogeneous depth)
  - Available software uses mostly simple, non-optimal recipes



Detecting galaxies

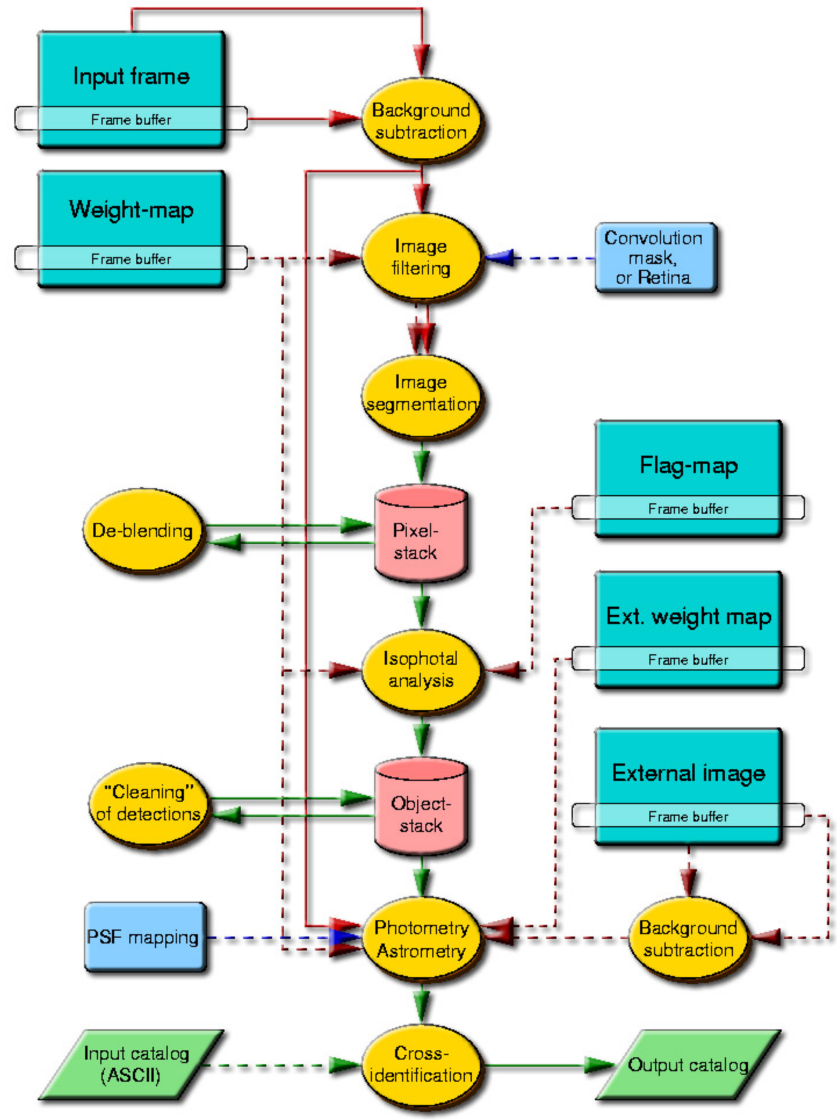
- To perform detection, one needs to make assumptions about what the sources look like
  - Point-sources:
    - use the Point Spread Function (PSF)
  - Galaxies and other diffuse objects appear in a wide range of shapes
    - The point-source definition is extended to “stuff that looks like a bright spot”
  - The wings of objects fade in the background noise, and overlap with other sources. How to define the object boundaries, and what is what?
    - Use isophotal limits (e.g. SExtractor)
    - Use a model of the profile

Providing enough computing power to process the data



- This does not take into account I/O bottlenecks
- Many current experiments require pipeline throughputs of at least ~1Mpix/s and ~ 100-1000 sources/s
- Parallelisation mandatory to keep up with the data rate

Source Extraction: SExtractor

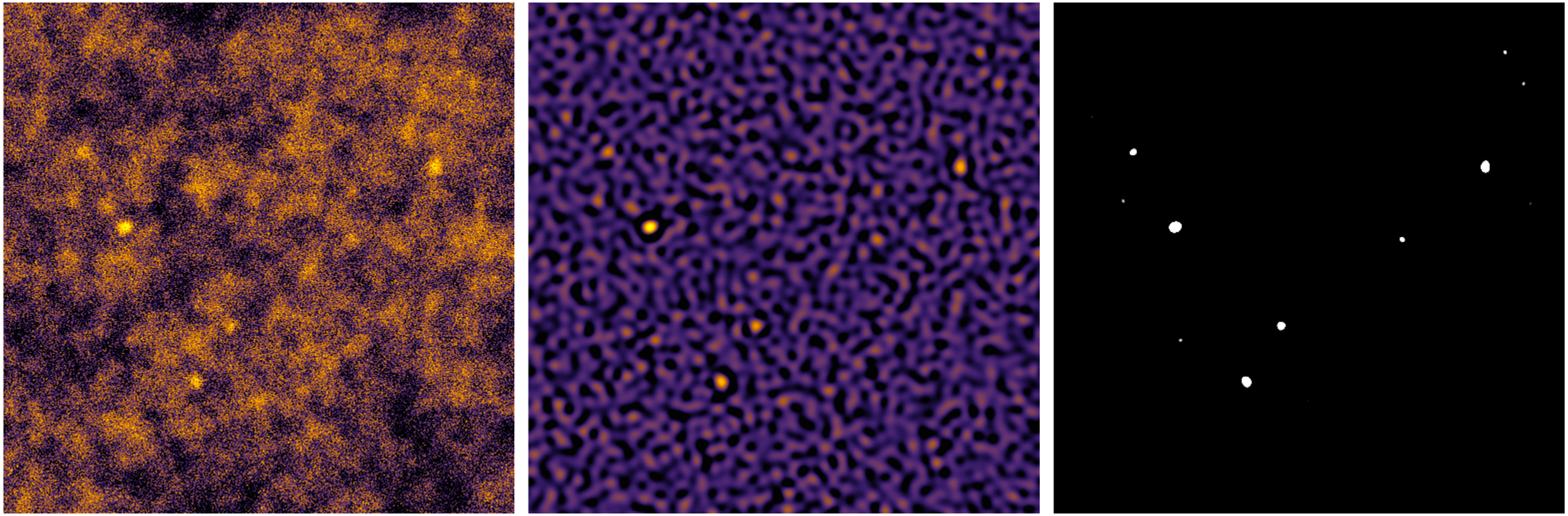


Basics of detection

- Broadly, the goal of detection is to discern between signal and noise.
  - In the sense of hypothesis testing, one wishes to provide a measurement whose value is a test that provides at every place in the image the best discrimination between “there is nothing there” and “there is a galaxy (or a star)”.
- Isolated case: apply some threshold to the image after increasing the contrast of sources with respect to the background noise.
- Linear filtering: for an <u>isolated</u> profile $\phi(\boldsymbol{x})$ superimposed to a (wide-sense stationary) background noise with spectral power $P(f)$, the optimum filter is the convolution with the *matched filter*

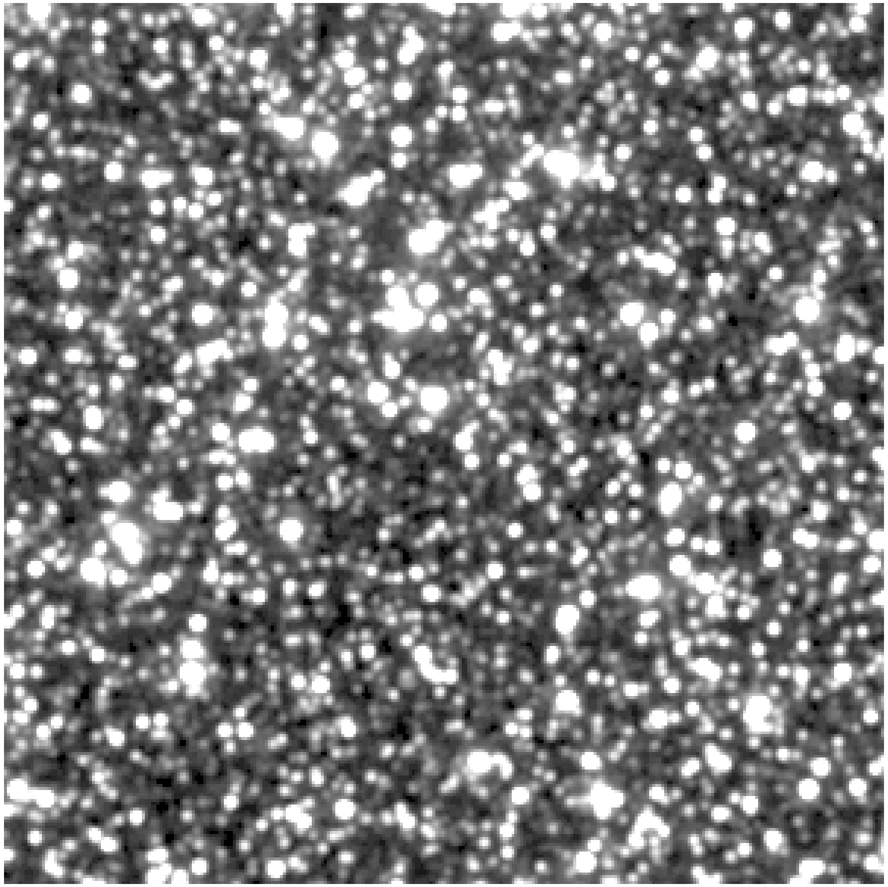
$$\boldsymbol{h} = \phi^T * \mathcal{F}\{P^{-1}\}$$

  - In many cases (unresampled CCD images with local background subtracted), the noise spectrum can be considered as “white” on source scales: $P(f) = cste$



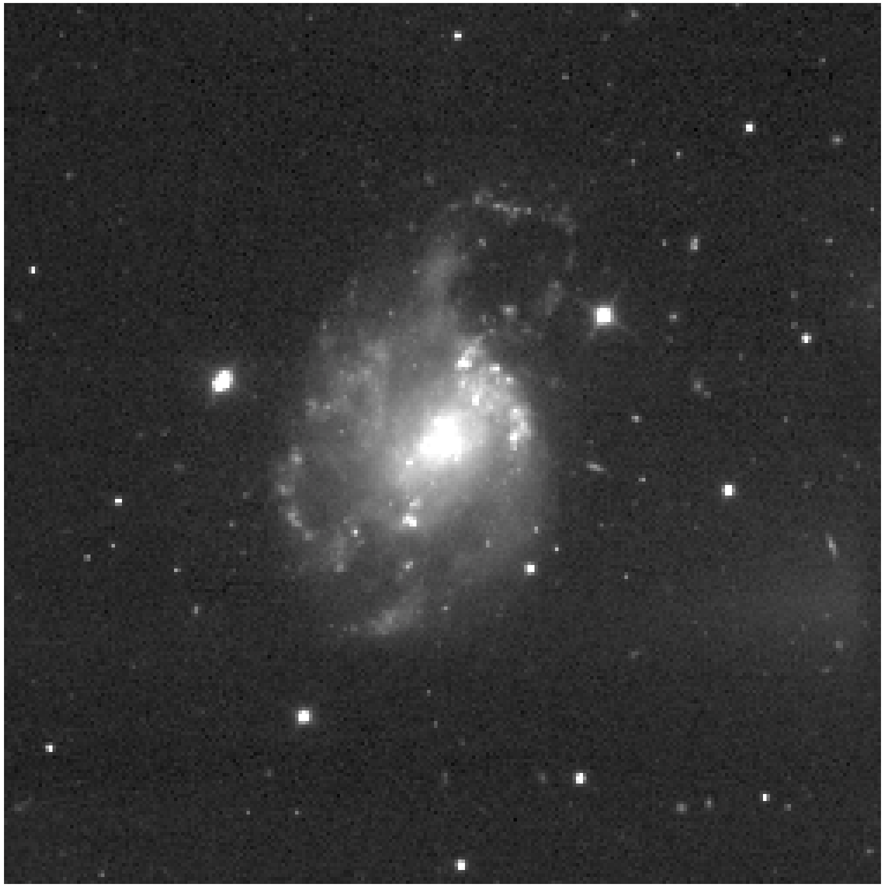
Astronomical sources are seldom isolated

- Confusion-noise limited fields of unresolved sources (e.g. crowded star fields)
  - The best linear filter becomes a deconvolution filter (for Poisson distributions)
  - 1st step: Detection of peaks
    - Correlator or basic peak-search
    - Efficient for crowded stellar fields
    - Quite unreliable for extended sources like patchy galaxy disks or low surface-brightness objects
  - 2nd step: Grouping of detections
    - Defines cluster of overlapping profiles which require simultaneous fitting
    - Such catalogs are limited in magnitude
    - DAOPhot (Stetson 1987), DoPhot (Schechter et al. 1993),… et plus généralement CLEAN (Hogborn 1974) ou encore MCS (Magain et al. 2006)



Galaxies occur in a variety of shapes

- Galaxy-oriented detection
  - 1st step: single-scale (PSF) or multi-scale filtering
    - Choice of a preferred scale.
  - 2nd step : Thresholding and segmentation
    - Efficient over a larger range of object scales
    - Threshold must be set low to detect faint objects, with the consequence that close sources are heavily blended
  - 3rd step: Deblending of detections
    - Can be done through local peak search, or multi-thresholding (more consistent)
    - Galaxy catalogs are surface-brightness limited for resolved objects.
    - FOCAS (Jarvis & Tyson 1981), Irwin 1985, SExtractor (Bertin & Arnouts 1996), Yoda (Drory 2003), Lupton 2005,…

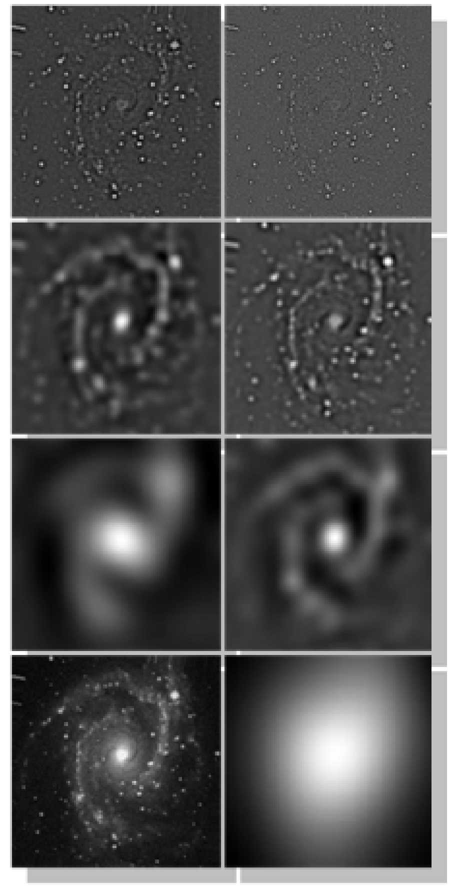


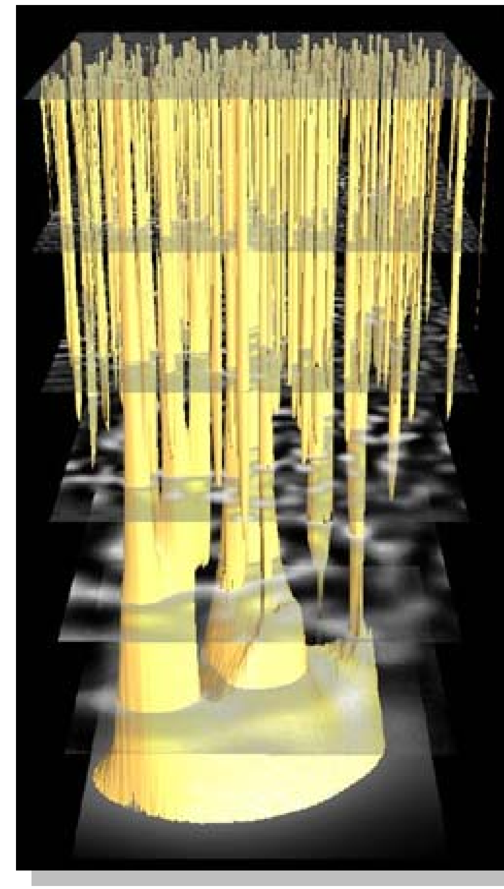
Multiscale analyses

- Extend the benefit of filtering from point-sources to very extended objects
  - Wavelet analysis: a data cube $w(\mathbf{x},a)$ is obtained by correlating the image with the basis functions

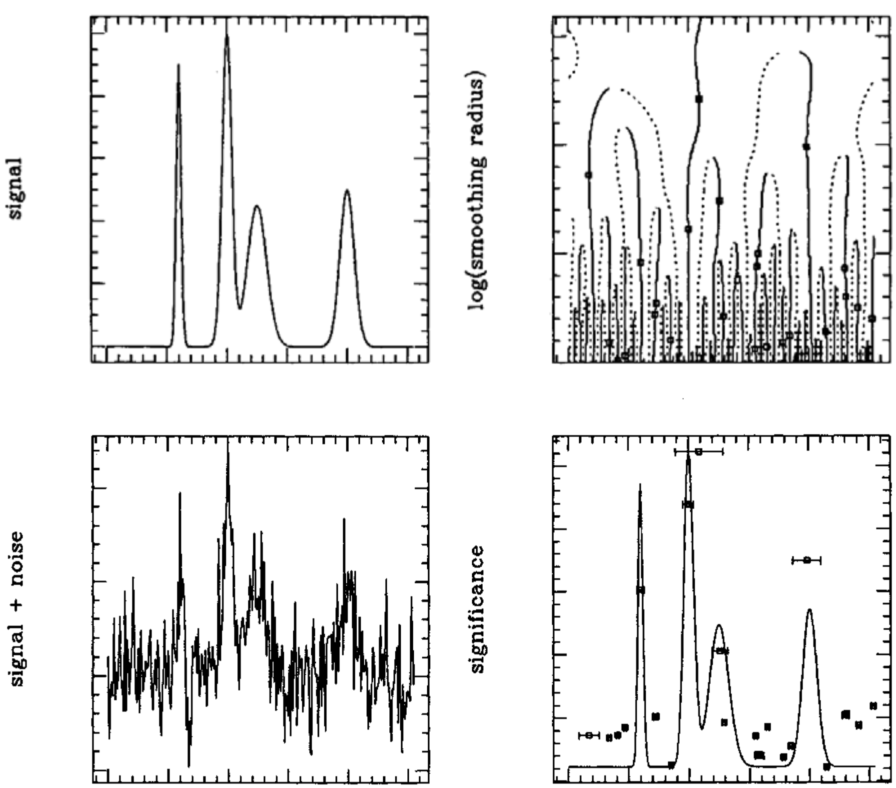
$$\psi_{a,\boldsymbol{b}}(\boldsymbol{x}) = \frac{1}{a}\psi\left(\frac{\boldsymbol{x}-\boldsymbol{b}}{a}\right)$$

  ψ is localised, isotropic, and has zero mean.
- Other multiscale analyses
  - Empirical method: set of band-pass filters (e.g. IMCAT, Kaiser et al. 1995)
  - Pyramidal median transform: linear decomposition using non-linear filtering (Starck et al. 1995)
- The last difficult (and not yet perfectly solved) step is to connect the detections done at each scale to reconstruct the final object (e.g. Bijaoui & Rué 1995).

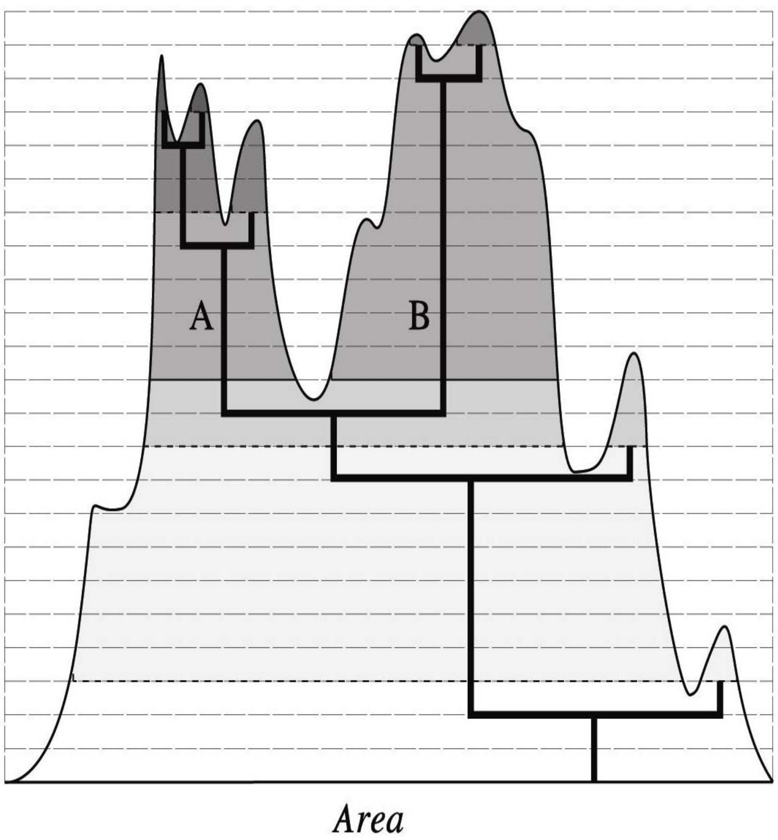




Multiscale / Multithreshold analysis



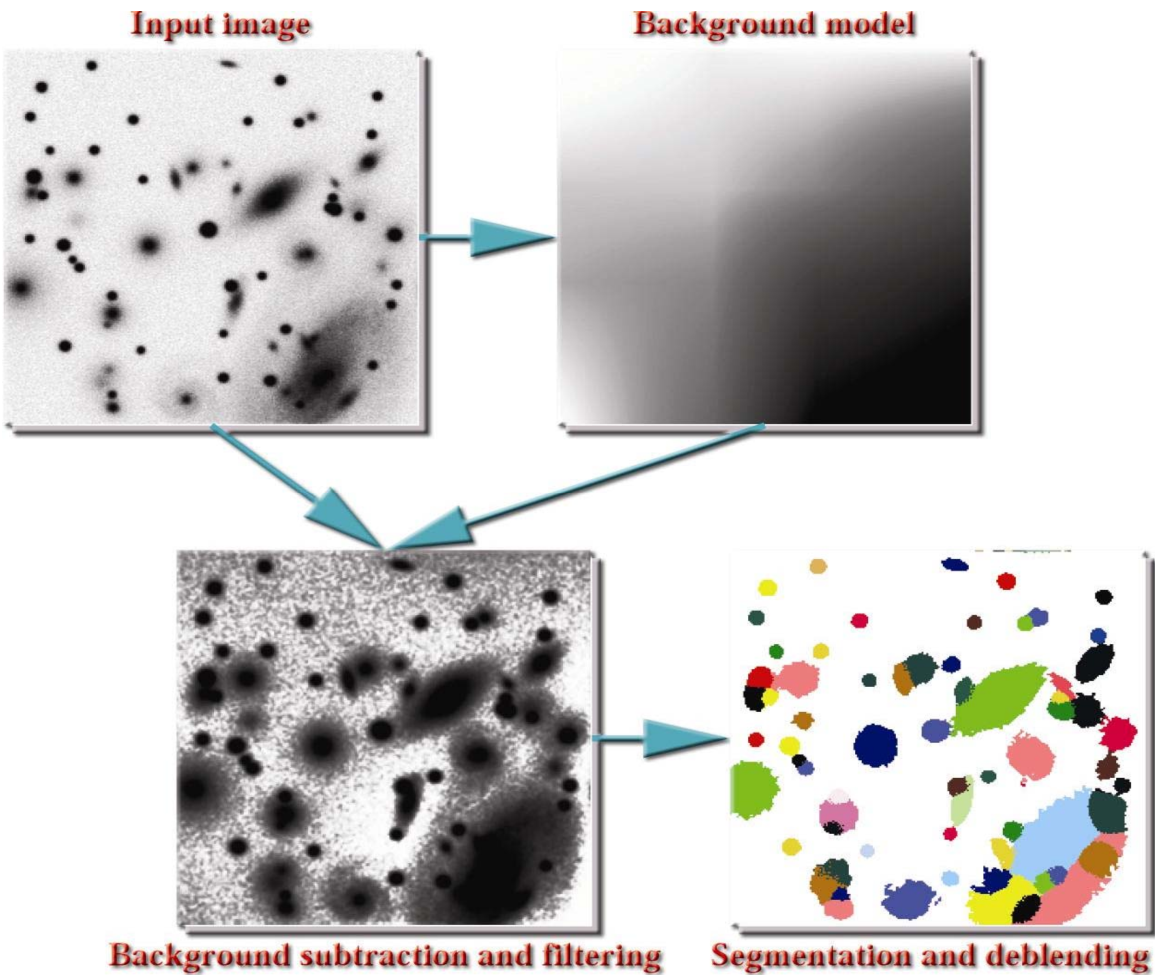
Kaiser et al. 1995



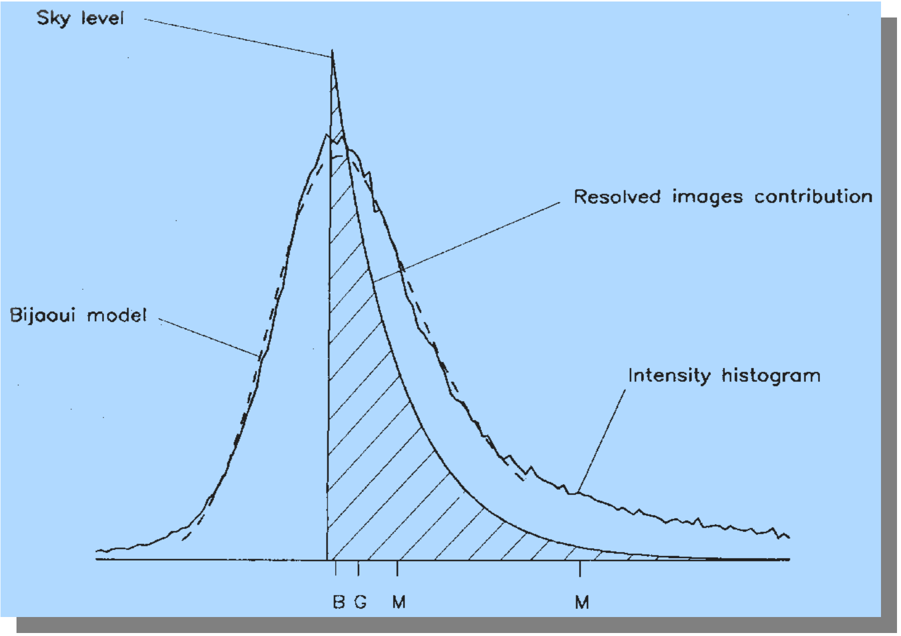
Bertin & Arnouts 1996

Operational source detection

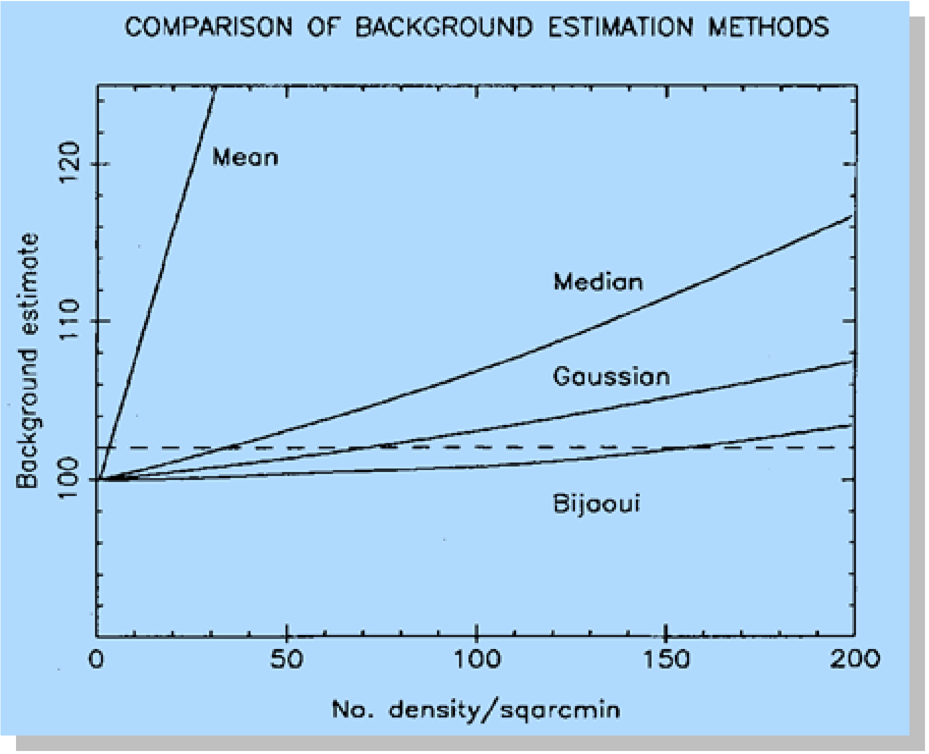
- The “traditional” approach involves
  - Sky background subtraction
  - Image filtering
    - Match filter using the PSF
  - Detection itself:
    - Local peak search, or
    - thresholding and image segmentation
  - Merging and/or splitting of detections
  - Exploration of surrounding pixels
  - “Cleaning” of spurious detections



Background estimation

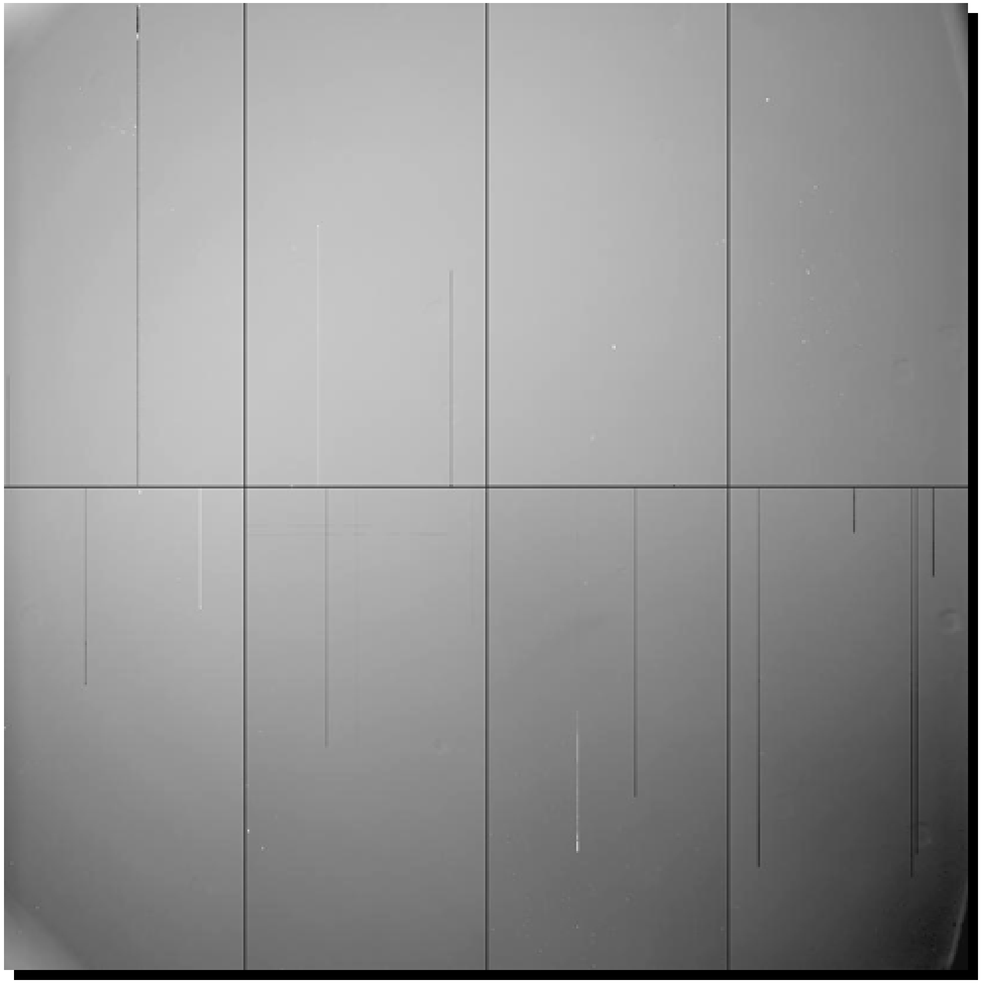


Irwin 1985



Weight-maps

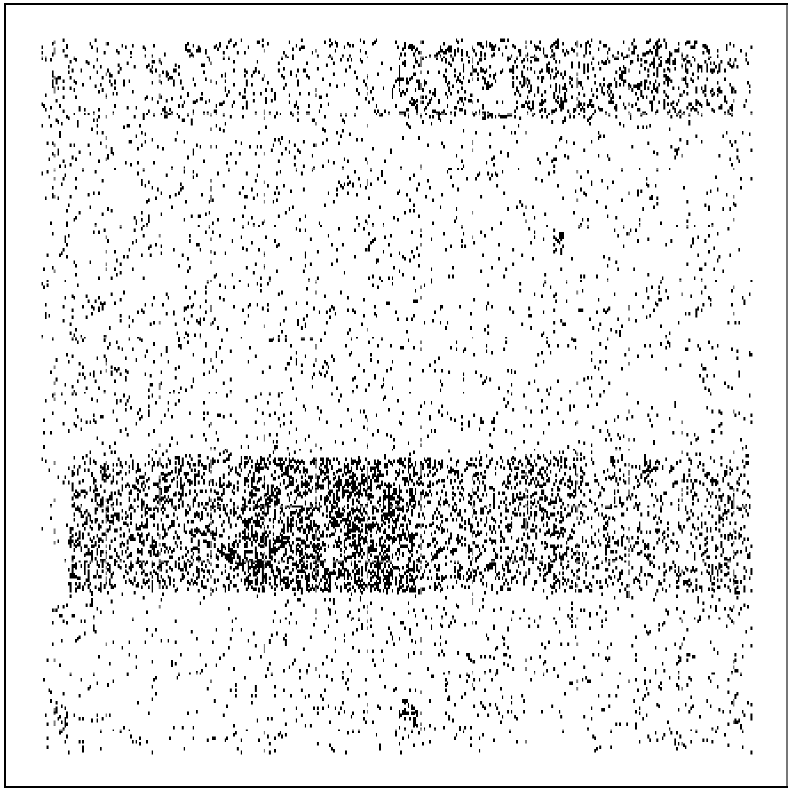
- Extend the "concept" of masking bad pixels:
  - Each pixel is given a weight $\propto 1/\sigma^2$
    - ➔pixel-to-pixel covariance is ignored.
    - ➔The photon-noise contribution from the sources themselves is ignored: images are supposed to be background-noise limited
      - Valid mostly for broadband imaging/large exposures
  - Image artifacts are given a weight of 0.



Using weight-maps for adaptive thresholding

$$t = k\sigma\sqrt{\sum_i \frac{h_i^2}{w_i}}$$

➜ Without weighting



➜ With weighting

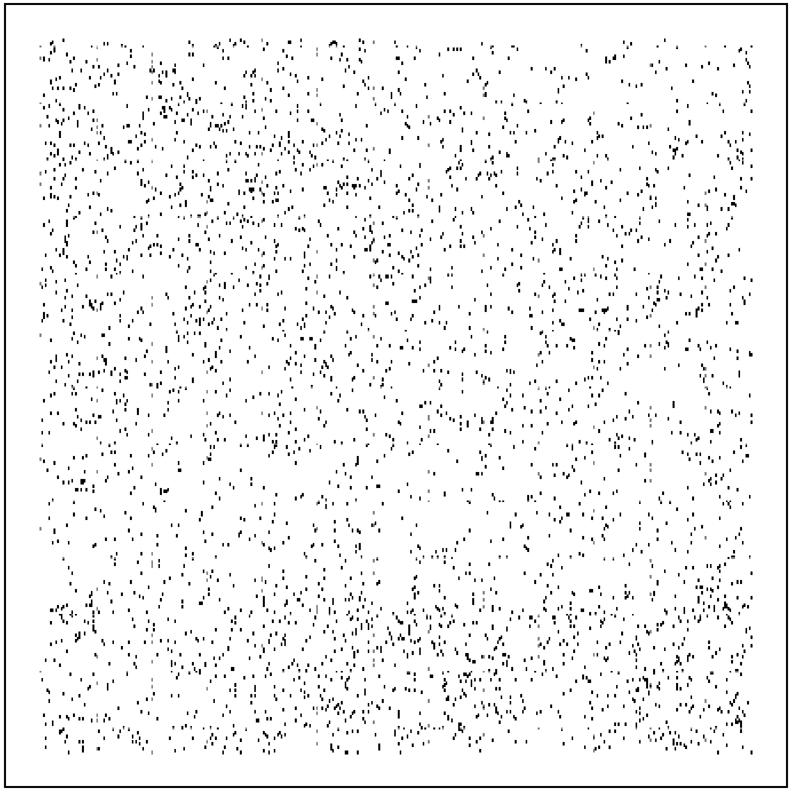
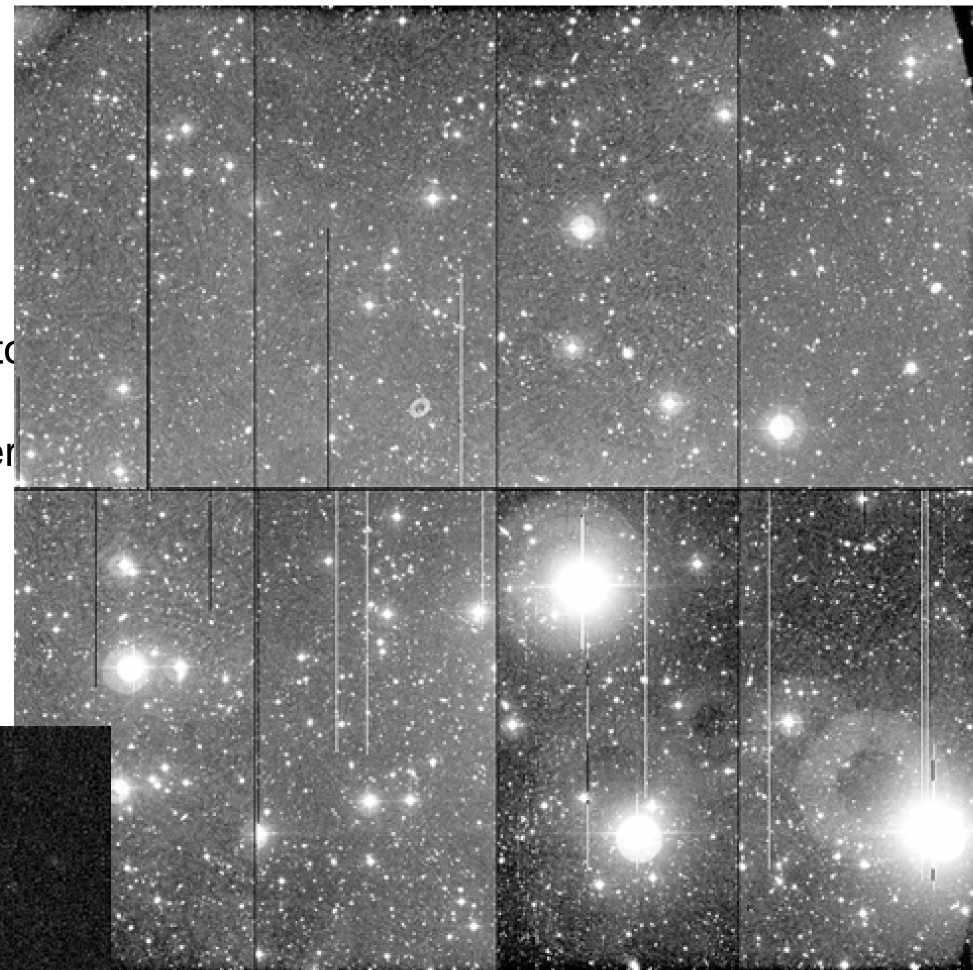


Image artifacts on wide-field exposures

- Cosmic ray hits
- Noisy/bad pixels / columns
- Spurious reflections and satellite trails are unavoidable on a 1 sq.degree field
  - Low surface-brightness halos due to reflections of bright stars in the refractive optics of the focal reducer
  - Diffraction spikes from the prime focus/secondary support
  - “Comet tails” close to the CCD borders



Bright stars are unavoidable in wide fields

- A handful of « annoying » stars at the galactic pole, 10 times more below *b*= 20 deg

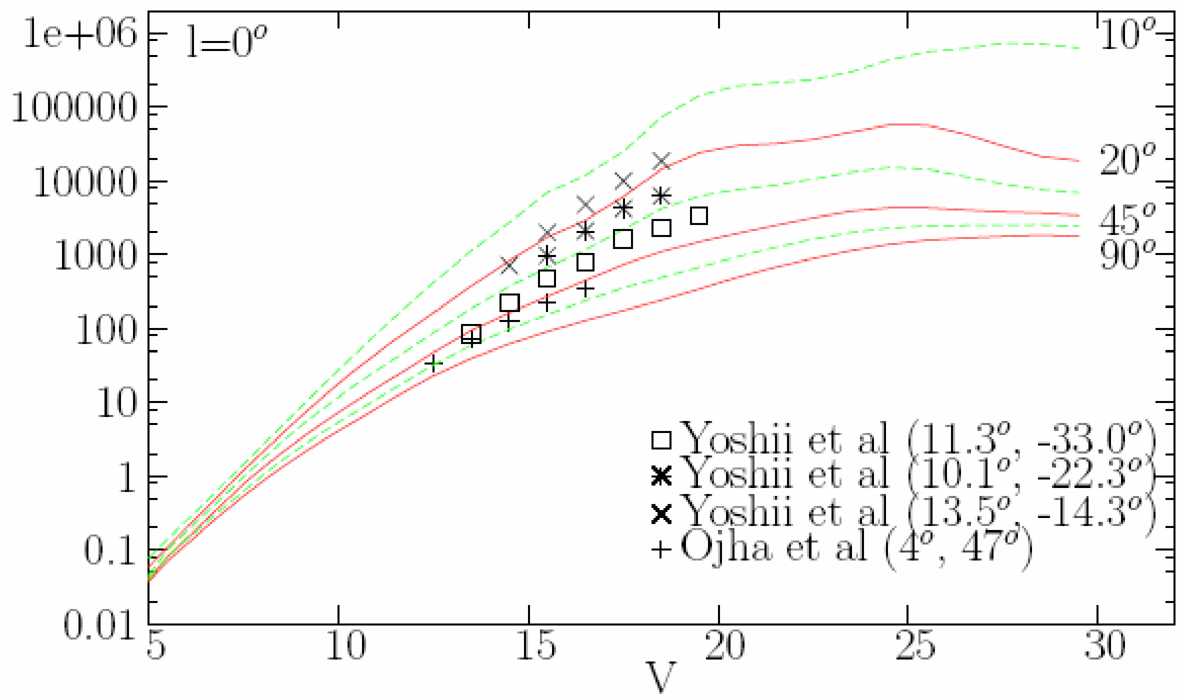
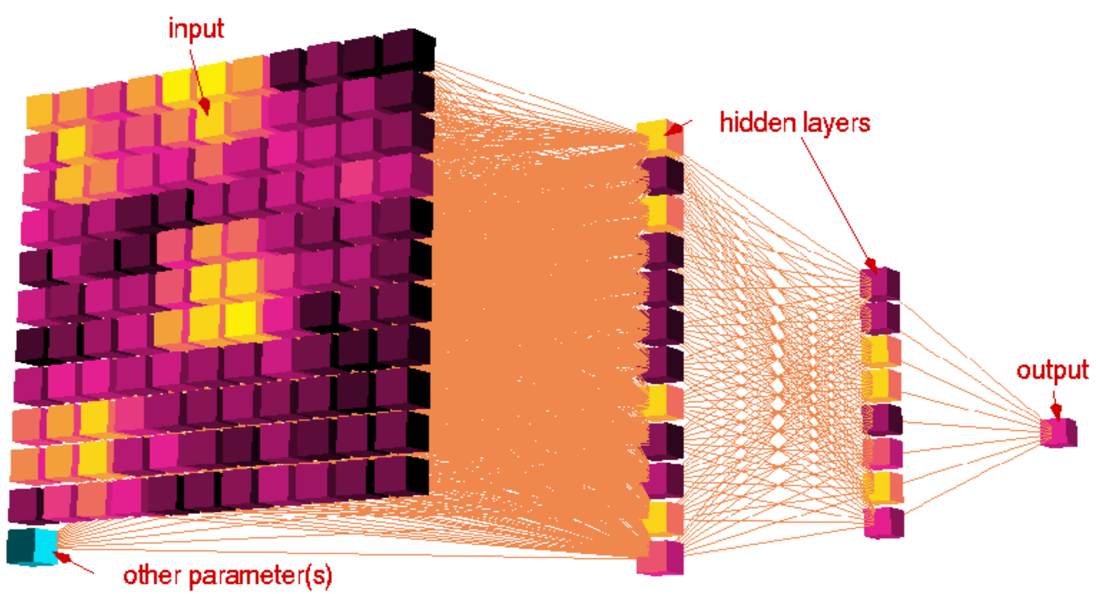


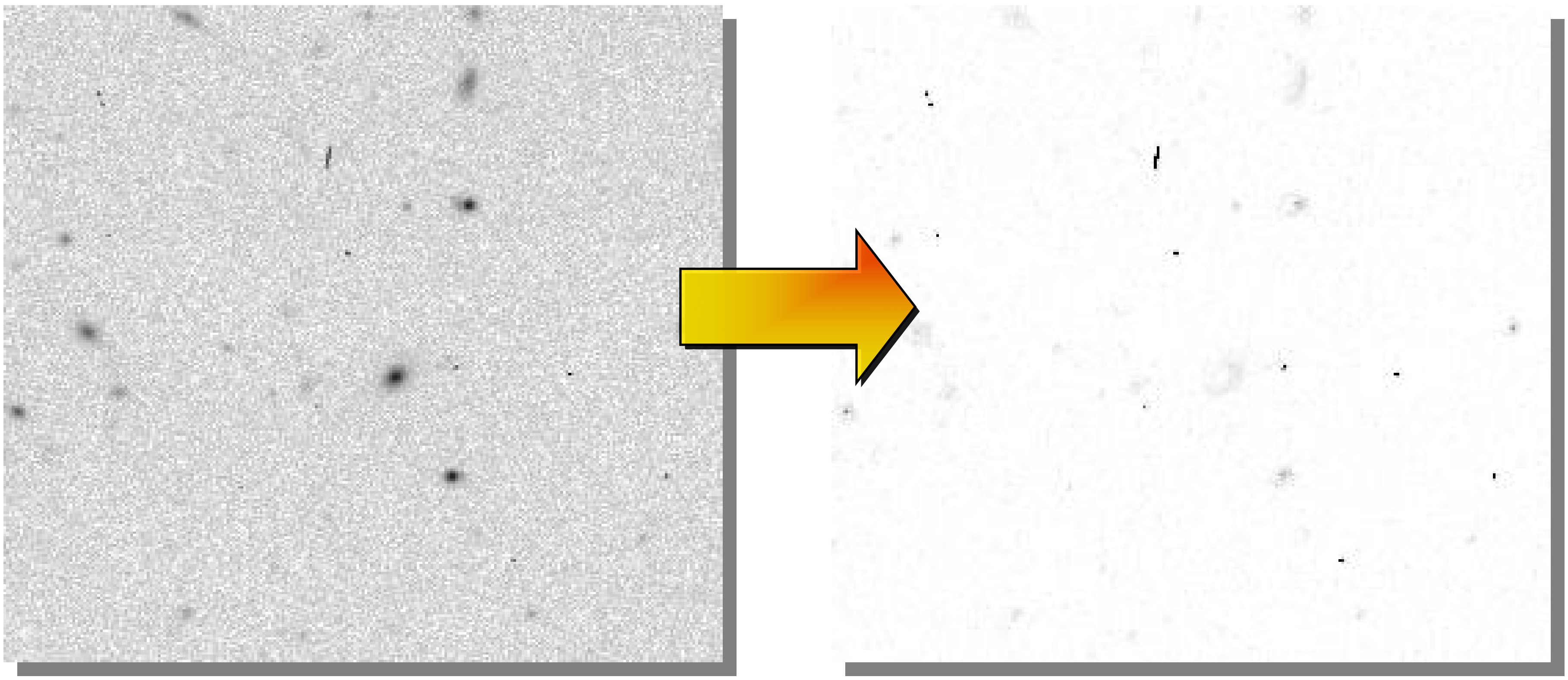
Fig. 5. Star count predictions (stars per magnitude and per square degree) in the V band at l=0°, for latitudes 10° to 90° from top to bottom (20°,45° and 90° with solid lines, 10°,30°,60° with dashed lines). Data are from Ojha et al. (1994a) and Yoshii & Rodgers (1989).

Automatic detection of small defects

- MLP directly connected to a 5x5 sliding mask applied to science exposures
  - "EyE" system (Bertin 1997)
  - Learning done on clean data + "dark exposures" containing localized defects
  - Dynamic range compression



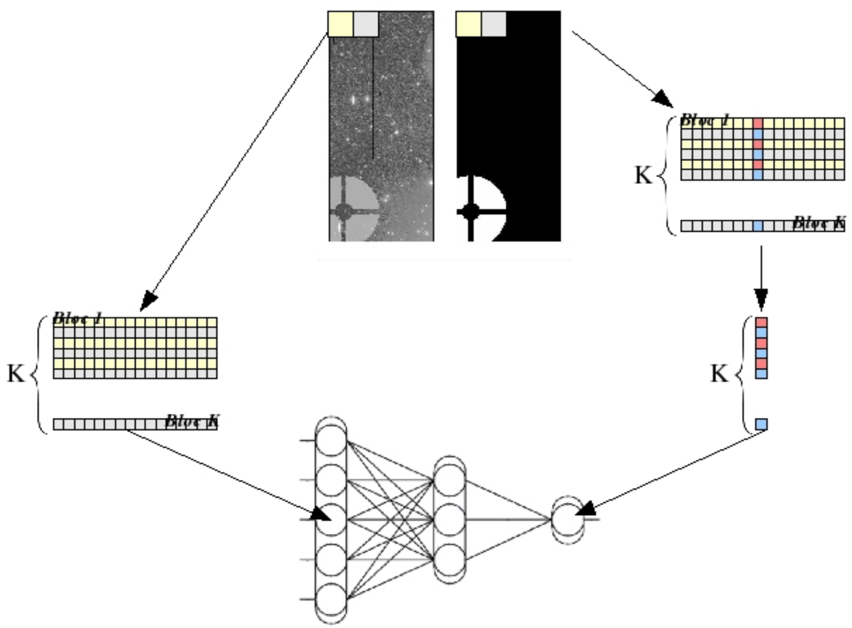
“Retina” filtering of small image defects



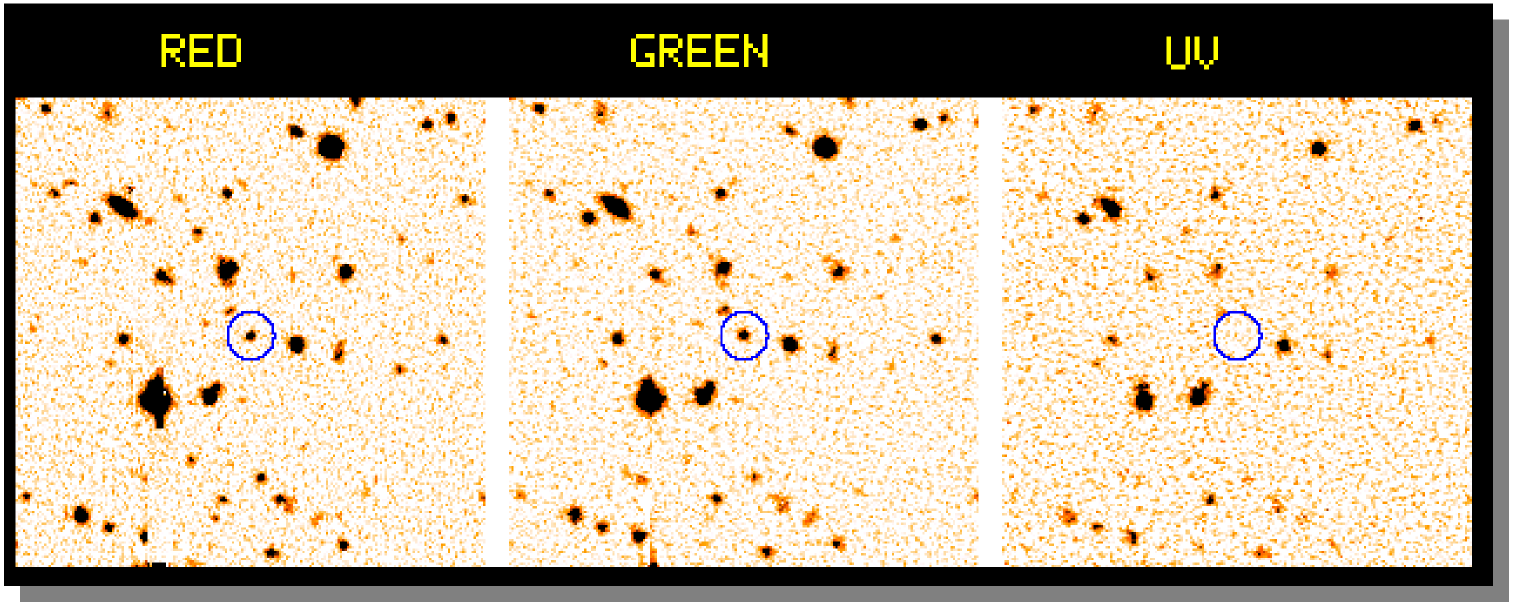
↑Input image **↑Output image**

Detecting extended features

- Identify extended image features like star halos and optical “ghosts” without triggering on galaxies (Baillard 2005)
- Dimensionality reduction layer (PCA)+MLP
- Analysis conducted at two scales



Panchromatic detection



← A "UV drop-out"

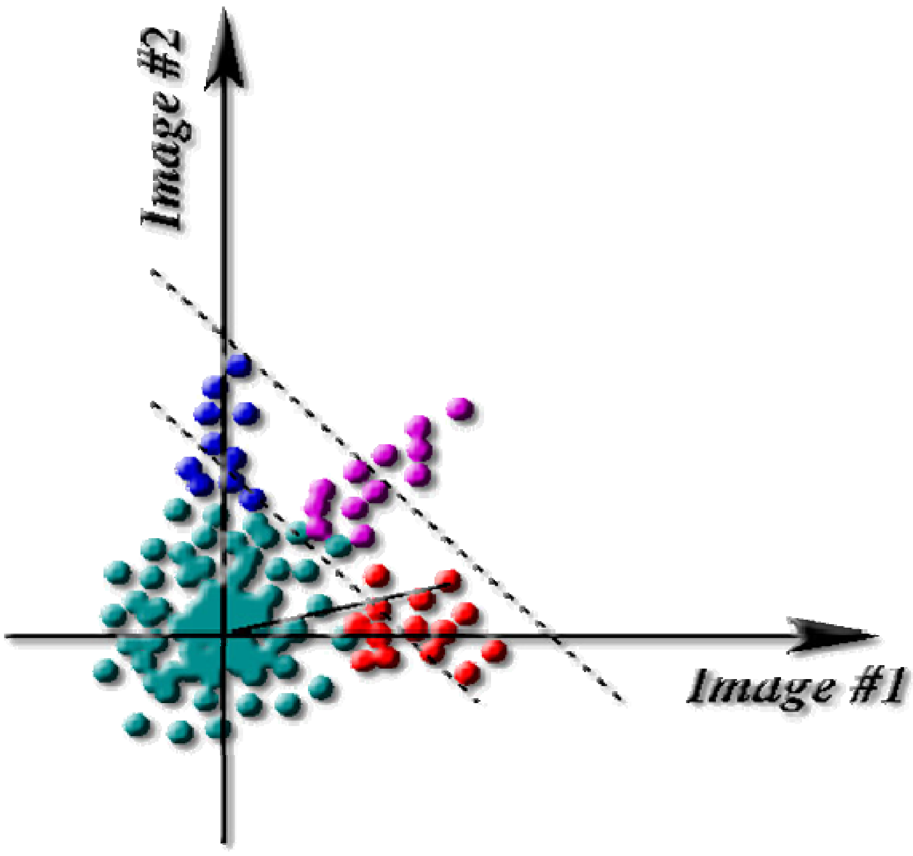
- In terms of signal-to-noise ratio, there exists an optimum *linear* combination of images of a source taken in different bands.
  - But it depends on the spectral properties of the source, and is therefore different for each object of an image.
- A non-linear combination exists, which is optimum for sky noise-limited images, regardless of the source spectrum: the "χ^2 image"

χ^2 images

- χ^2-like combination of the individual images (Szalay et al. 1999):

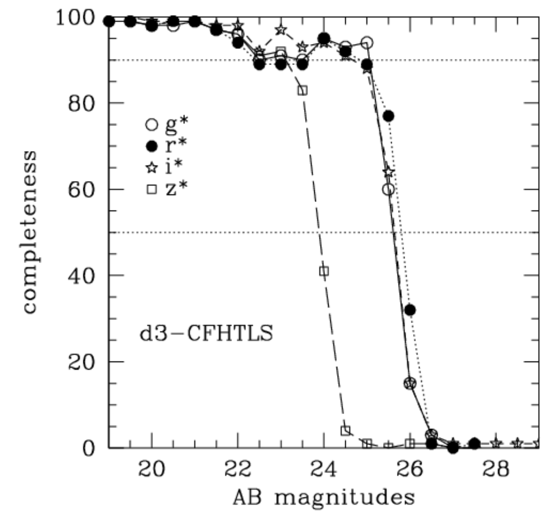
$$F = \frac{1}{n}\sqrt{\sum_i \frac{f_i^{\,2}}{\sigma_i^{\,2}}}$$

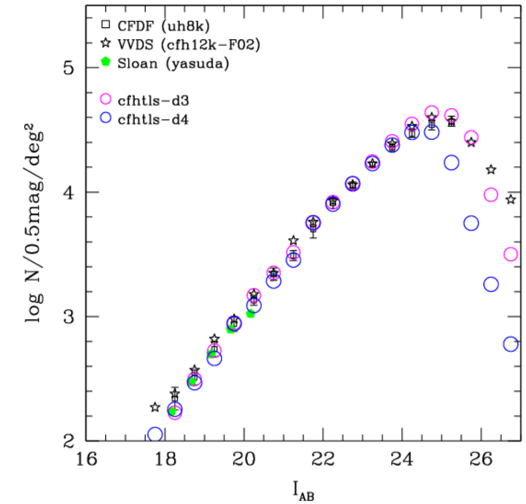
- Makes direct use of weight-maps
- The resulting image has a strongly non-linear response to flux
  - Can only be used as a "detection" image



Assessing the detection performance

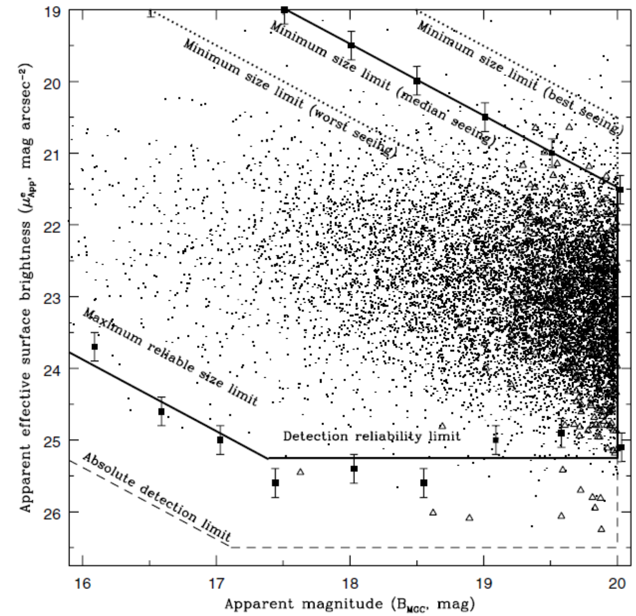
- Monte-Carlo simulations
  - Completeness check using mock sources added to the real images
  - Reliability checks:
    - running the detection algorithm on a negative version of the image
      - assumes a symmetric noise distribution
    - simulating a « realistic » survey image empty of sources
- Differential number counts
  - model-dependent
- Two-point correlation functions
  - may reveal local, spurious detection holes or clumps

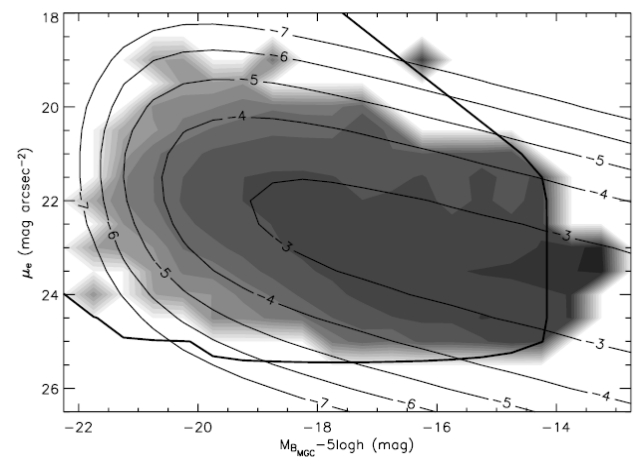




Detection reliability domain

- Surface brightness limits (Driver et al. 2005)
  - faint-end: detection threshold
    - may generate a bias against face-on galaxies in some infrared surveys
  - bright-end: detector saturation
- Size limit
  - small-end: point-source/resolution threshold
    - often reached in ground-based faint galaxy surveys
  - large-end: background modeling scale
- Flux limit
  - faint-limit: background-noise limit
- Environment
  - The two-point correlation function must vanish at separations < galaxy size
  - Faint objects cannot be detected too close to bright ones





Flux measurements for galaxies

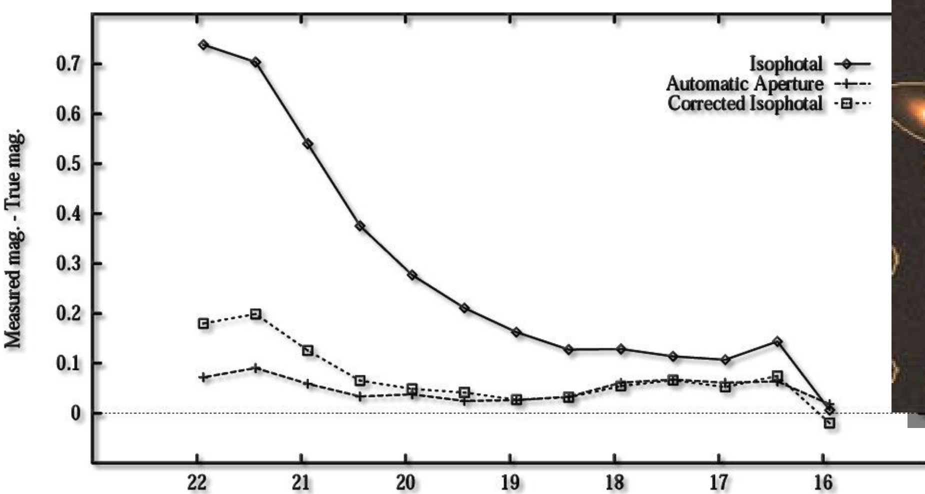
- How to measure the “total” flux of a fuzzy object with unknown shape?
  - Isophotal magnitudes: pixel values are integrated within a given isophote
    - Fast and simple
    - Consistent with the idea that sources are uniquely defined by a list of pixels
    - Reasonably robust to contamination by close neighbours
    - Fairly efficient in terms of signal-to-noise
    - Strongly biased against faint and low-surface brightness sources
    - Unless the limiting isophote is very low, or a Monte-Carlo model is available for comparison, should be used for rough magnitude estimates only
  - Aperture magnitudes: pixel values are integrated within a circular aperture
    - Fast and simple
    - Unbiased against faint or low surface brightness sources
    - Contamination by close neighbours can be strong
    - Rather inefficient in terms of S/N
    - Can be used whenever the data are meant to be compared with external measurements
      - Photometric calibration of standard stars (e.g. Landolt)
      - Colour measurements
  - Adaptive aperture magnitudes: pixel values are integrated within a circular or elliptical aperture which is automatically scaled to the object
    - Needs 2 passes through the data
    - Weakly biased against faint or low surface brightness sources
    - Contamination by close neighbours must be dealt with
    - Fairly efficient in terms of S/N
    - Supposed to provide a kind of “all ground photometry” with typical accuracy $\approx$ 0.1 mag.
      - Large galaxy samples
      - OK for stars at high galactic latitude
      - Caution needed for low surface brightness stuff

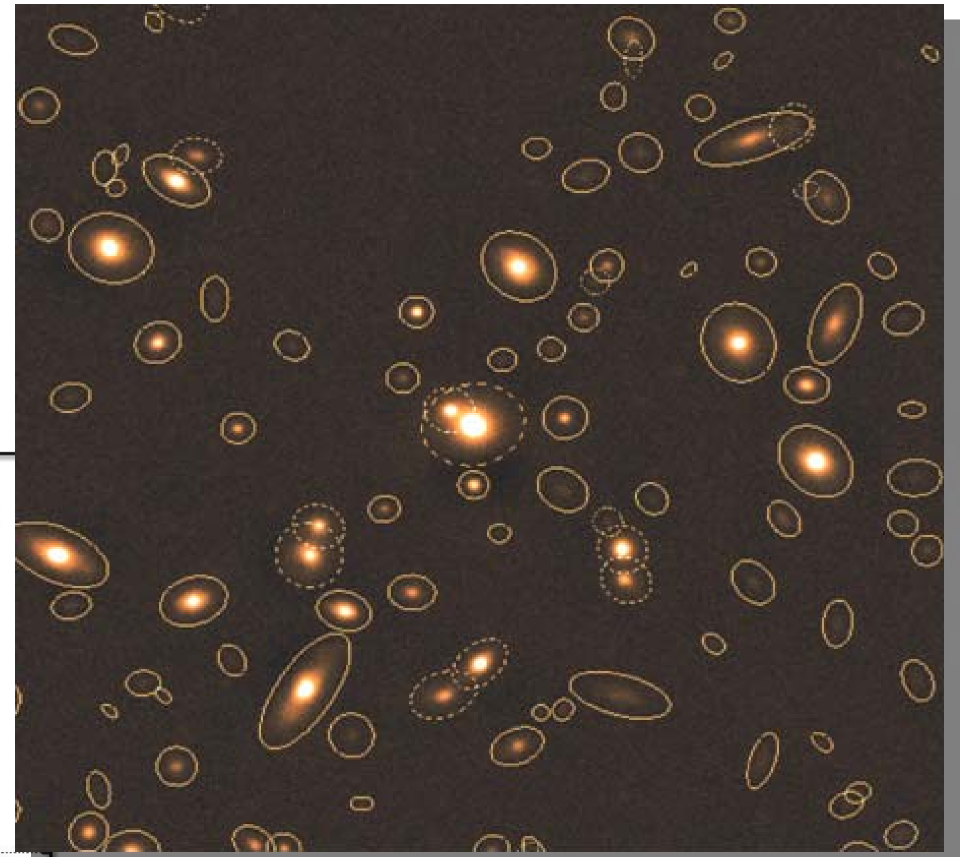
Adaptive aperture photometry

- Kron magnitudes (Kron 1980)
  - Scale the aperture with the "1st order radial moment" r_1 :

$$r_{\mathrm{lim}} = k.\underbrace{\frac{\sum rI(r)}{\sum I(r)}}_{r_1}$$

  - Efficient and "surprisingly" robust, even for faint objects





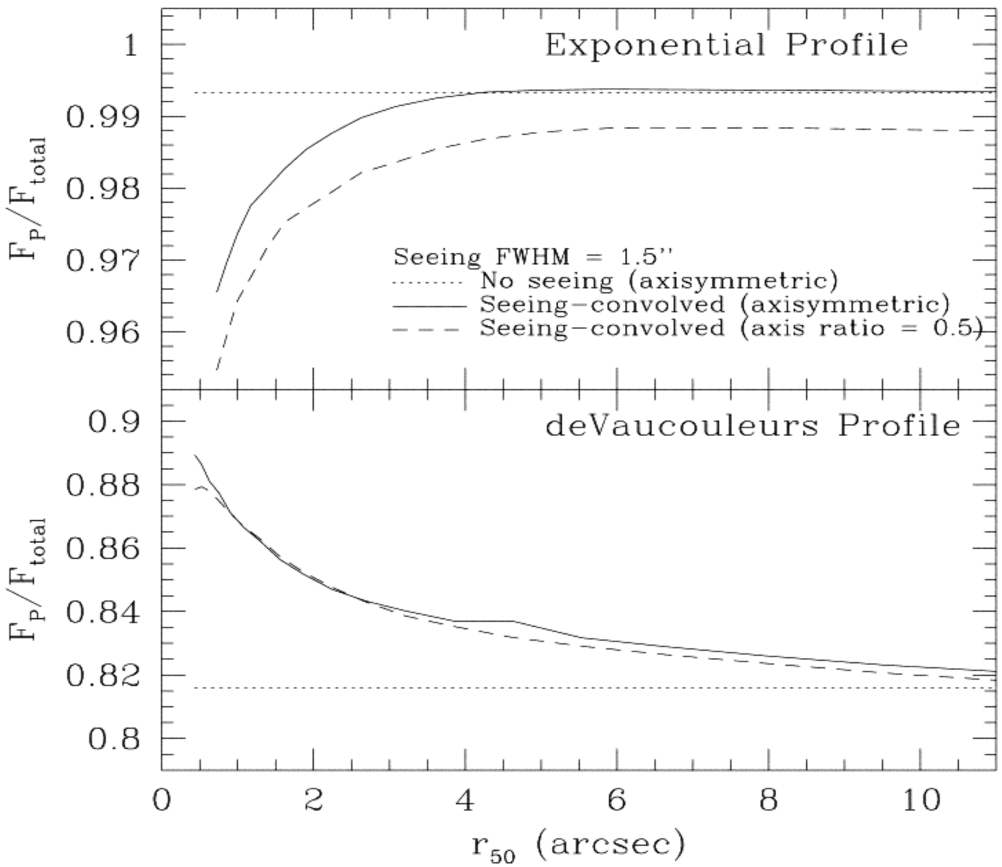
Adaptive aperture photometry

- Petrosian magnitudes (Petrosian 1976)
  - Find the radius where the local surface brightness is a given fraction of the average surface brightness within the enclosed disk, and use it to scale the aperture.

$$r_{\mathrm{lim}} = N_{\mathrm{P}}.r_{\mathrm{P}}$$

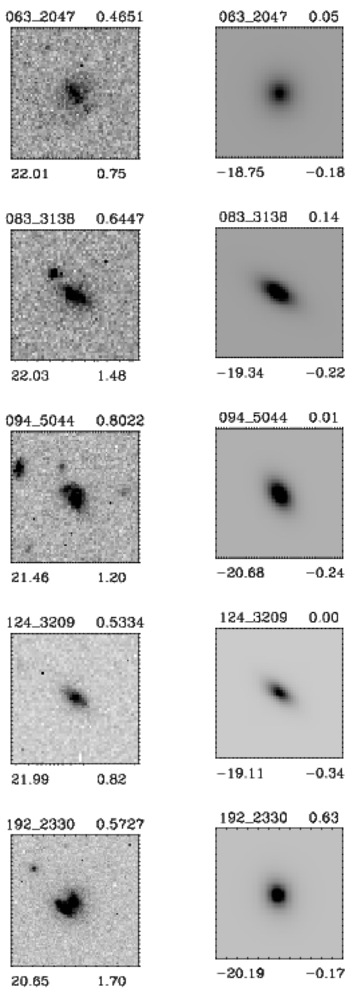
$$\exists R_{\mathrm{P}}(r_{\mathrm{P}}) = R_{\mathrm{P,lim}}$$

$$R_{\mathrm{P}}(r) = \frac{\sum_{\alpha_1 r < r' < \alpha_2 r} I(r')/((\alpha_2^2 - \alpha_1^2) r^2)}{\sum_{r' < r} I(r')/r^2}$$



  - Used by SDSS (e.g. Blanton et al. 2001)
  - The most accurate for resolved galaxies with good S/N
  - For low S/N, performance slightly worse than Kron magnitudes

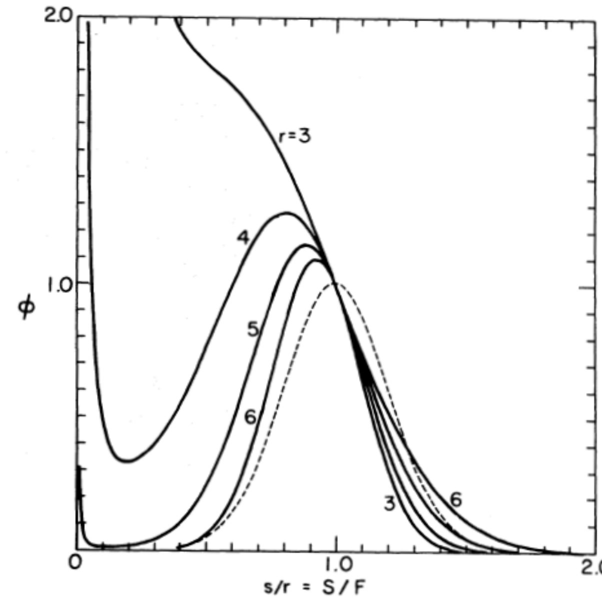
“Asymptotic” profile-fitting photometry

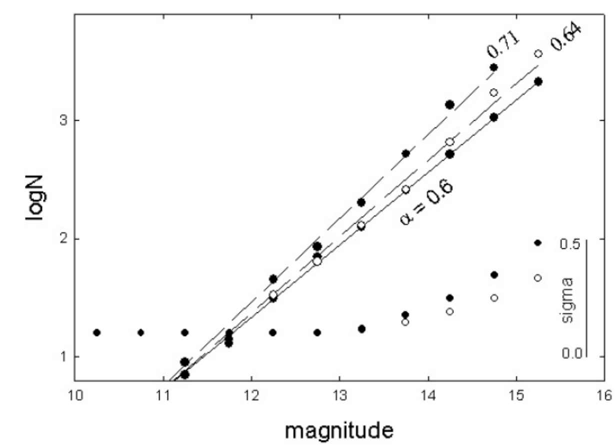


- Parametric decomposition: bulge+disk or Sersic
  - Works at very low S/N per pixel
  - The model is reconvolved by the PSF at each iteration
  - Metropolis-type algorithm can be necessary to escape local minima
  - High computational cost
  - Asymmetry and spiral arms introduce noise
  - Appears to give more reliable measurements (`cmodel`) than Petrosian magnitudes in SDSS

Photometric biases

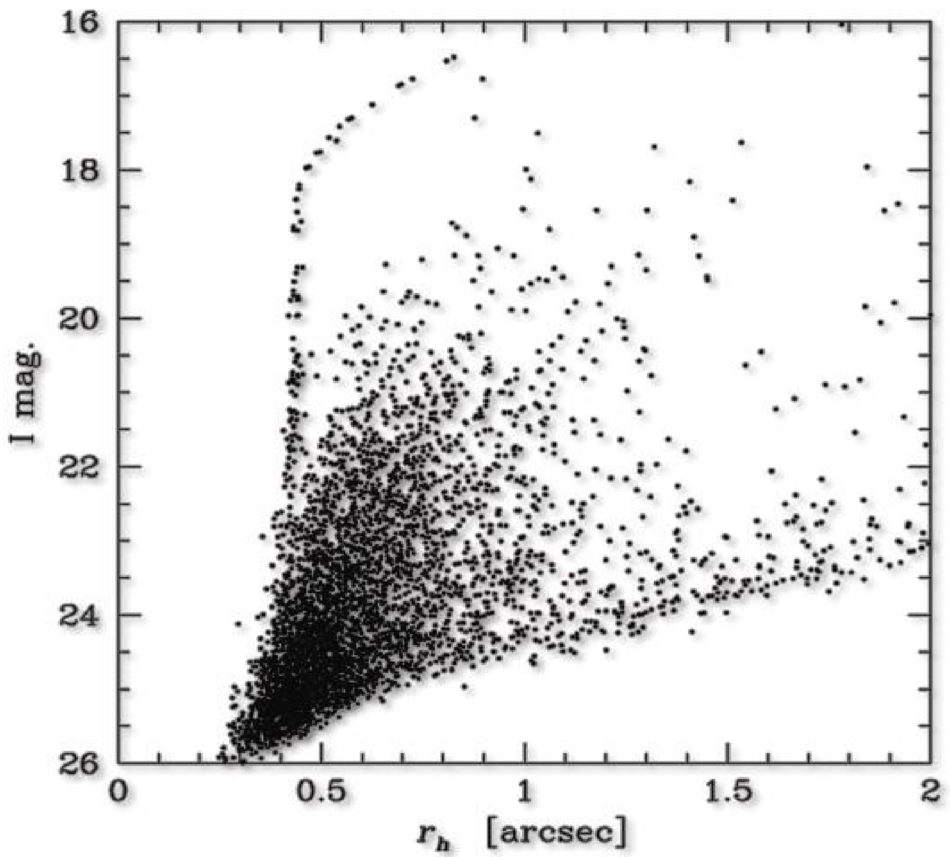
- Type-related
  - The wings of early type galaxies are more "shallow" than those of late types
    - higher Sersic index *n*
    - a larger flux fraction may be missed (up to ~40%)
    - *k*-corrections make spheroids vanish in the optical at high redshift
- Environment-related
  - Profile overlaps in dense galaxy clusters
- Noise-related
  - Eddington bias (1913): the strong 2nd derivative of differential galaxy number counts artificially boost the counts, especially above the completeness limit (mostly unresolved sources)
    - Close to the noise limit, detected sources stand preferably on positive noise peaks
    - Below 5σ accurate Monte-Carlo simulations may be needed to correct for this bias (Murdoch et al. 1973, see also Teerikorpi 2004)





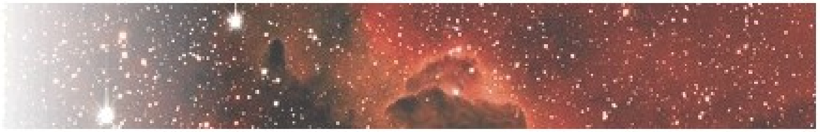
Automatic star/galaxy classification

- Mandatory for deep imaging surveys at high galactic latitude. Number density of galaxies = number density of stars at V~20 at high galactic latitude
- In the optical domain: based on shape
  - Multi-dimensional analysis in shape parameter space
  - Priors concerning the relative number of objects at a given magnitude must be taken into account



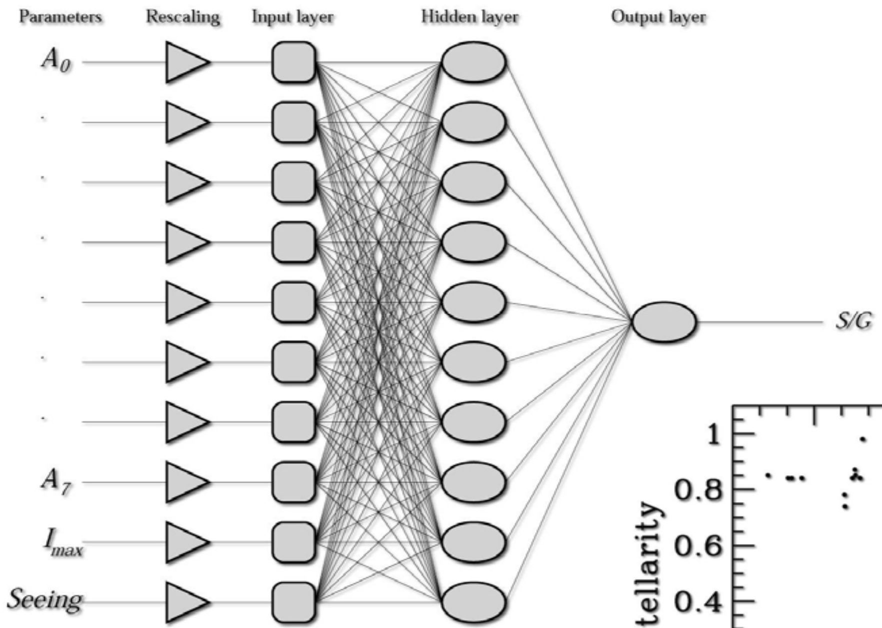
Automatic classifiers

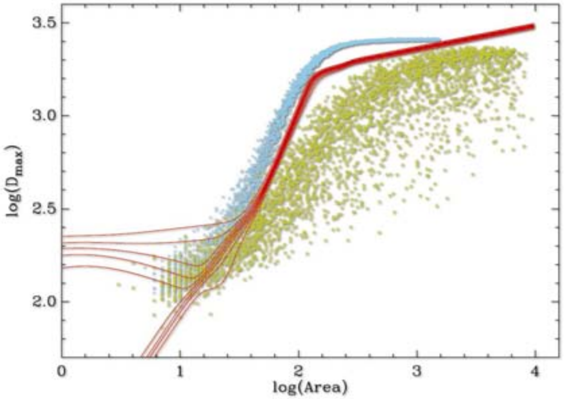
- FOCAS: Bayesian, based on a simple PSF model assuming that extended objects have the same profile as the PSF, but with larger FWHM.
- SExtractor: Artificial neural network trained on simulated ground-based images

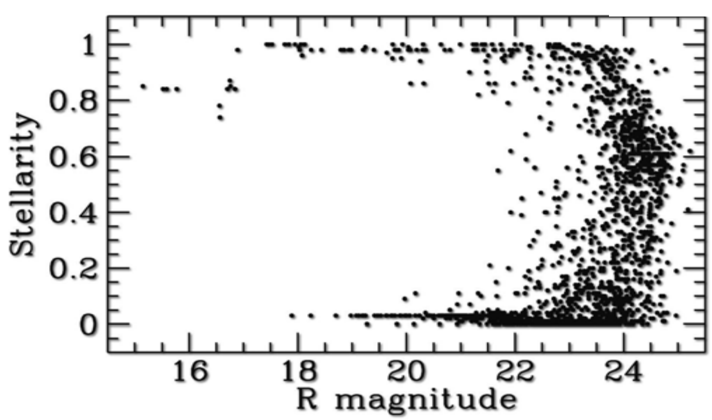


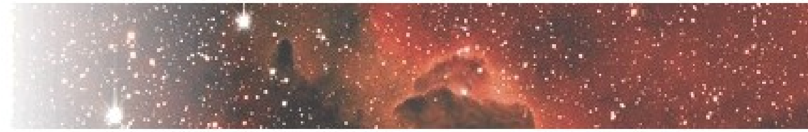
Star/galaxy classification

- SExtractor's CLASS_STAR is the output of an artificial neural network trained on simulated ground-based images
- One of the inputs acts as a "tuning button" set to the current PSF FWHM ("seeing")









terapix.iap.fr