Where CWI touches the Sky

Bart Scheers

Centrum Wiskunde & Informatica, Amsterdam
Astronomical Institute "Anton Pannekoek", University of Amsterdam

February 3rd, 2011
The "Field" and its unpaved Paths...
From Galaxies to Stars to (Exo)Planets

Supernova Remnants | Gamma-Ray Burst Afterglows | Exoplanets

Accreting Binaries | Radio Galaxies
Common Design Strategies

- Search for rapid transients and variable sources
- High-speed, all-sky surveys
- Cataloguing all measurements
  - Times Series, Lightcurves, Sky & Source Models
  - Make available for data mining, discovery
Next-generation Telescopes

**Common Design Strategies**
- Search for rapid transients and variable sources
- High-speed, all-sky surveys
- Cataloguing all measurements
  - Times Series, Lightcurves, Sky & Source Models
  - Make available for data mining, discovery

**Common Data "Challenges"**
- Lots of data are produced, and need to be processed and stored in near real time
- Even more, we want to access it all the time, with rapid response times
Next-generation Telescopes

**Common Design Strategies**
- Search for rapid transients and variable sources
- High-speed, all-sky surveys
- Cataloguing all measurements
  - Times Series, Lightcurves, Sky & Source Models
  - Make available for data mining, discovery

**Common Data ”Challenges”**
- Lots of data are produced, and need to be processed and stored in near real time
- Even more, we want to access it all the time, with rapid response times
- Databases...?
Common Design Strategies

- Search for rapid transients and variable sources
- High-speed, all-sky surveys
- Cataloguing all measurements
  - Times Series, Lightcurves, Sky & Source Models
  - Make available for data mining, discovery

Common Data "Challenges"

- Lots of data are produced, and need to be processed and stored in near real time
- Even more, we want to access it all the time, with rapid response times
- Databases...?
  - MonetDB maybe...?
LOFAR, the Low-Frequency Array

Where CWI touches the Sky
LOFAR, the Low-Frequency Array
Data Production: Expected Volumes & Rates

- Raw data \( \sim 25 \text{ TB/hr} \)
- Distinct sources: \( \sim 10^7 - 10^8 \),
  - which are revisited many, many, many times
- Source properties reduce to 50 – 100 TB/yr
- Peaks over 10,000 sources per second
Data Production: Expected Volumes & Rates

- Raw data $\sim 25$ TB/hr
- Distinct sources: $\sim 10^7 - 10^8$,
  - which are revisited many, many, many times
- Source properties reduce to $50 - 100$ TB/yr
- Peaks over 10,000 sources per second

(Near Real-Time) Data Processing

- Automated Software Pipelines
  - Imaging Pipeline
  - Transients Pipeline
- Fast data access, Quick responses
- Maintain Statistical Model of Sky
- Spread data
Dimensionless distance takes errors into account

\[ r_{ij} = \sqrt{\frac{(\alpha_i \cos \delta_i - \alpha_j \cos \delta_j)^2}{\sigma_{\alpha_i}^2 + \sigma_{\alpha_j}^2} + \frac{(\delta_i - \delta_j)^2}{\sigma_{\delta_i}^2 + \sigma_{\delta_j}^2}} < r_{\text{lim}} \]
Rayleigh Distribution: probability of finding source at $r \geq \rho$

$$p(r \geq \rho) = \exp(-\rho^2/2)$$
Time Series Data: Monitoring Source Variability in Lightcurves

Based on reduced $\chi^2$ statistics

$$\eta_\nu = \frac{N}{N-1} \left( \frac{w_\nu I_\nu^2}{w_\nu} - \frac{w_\nu I_\nu^2}{w_\nu^2} \right)$$
Time Series Data: Monitoring Source Variability in Lightcurves

$H_0$: source is not variable

Reduced $\chi^2$ probability justifies rejection/acception of $H_0$.

$$p_{\eta_\nu} = \int_{\eta_{\nu}'}=\eta_\nu \ p_{\eta_\nu}(\eta_\nu', N - 1) d\eta_\nu'$$
Time Series Data:
Monitoring Source Variability in Lightcurves

![Graph showing variability in lightcurves over time.](image-url)
We want to compare our observations

- with renown external static catalogues
- with our own dynamic catalogue, i.e. the Sky Model
We want to compare our observations

- with renown external static catalogues
- with our own dynamic catalogue, i.e. the Sky Model

During observations, we maintain statistical parameters

- $N$
- $I_\nu$
- $I_\nu^2$
- $w_\nu$
- $w_\nu I_\nu$
- $w_\nu I_\nu^2$

- plus position & errors, ...
We want to compare our observations
  - with renown external **static** catalogues
  - with our own **dynamic** catalogue, i.e. the Sky Model

During observations, we maintain statistical parameters
  - \( N \)
  - \( \bar{I}_\nu \)
  - \( \bar{I}^2_\nu \)
  - \( \bar{w}_\nu \)
  - \( \bar{w}_\nu I_\nu \)
  - \( \bar{w}_\nu I^2_\nu \)
  - plus position & errors, ...

But, we want more monitoring indices **inside** the database
  - Extend SQL: SciQL
Inspired by SciLens:
Sharded Lightcurve Database (50 – 100 TB/yr)
Conclusions & Open Issues

- Computer science, informatics & database architecting enter astronomy
- A good statistical model of the sky alleviates imaging and predicts "Sky Weather"
- SciQL will advance data mining
- Sharded database reduces data replication
- Move detection algorithms more upstream