Science Drivers for Big Data

Joseph Lazio
SKA Program Development Office &
Jet Propulsion Laboratory, California Institute of Technology

Data Intensive Astronomy

- “There is nothing new under the Sun ….”
- Case studies
  1. Processing: Cosmology and imaging surveys
  2. Data volume: Fundamental physics from pulsar observations and pulsar surveys
  3. Data rates: Observing!
  4. Data visualization: Identifying interference
  5. Data curation: Astronomy for the future

Data Intensive Astronomy

Data Volumes

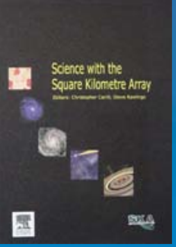
Ιππαρχο∫ (Hipparcus)

- ca. 135 BCE
- Stellar catalog with 850 entries

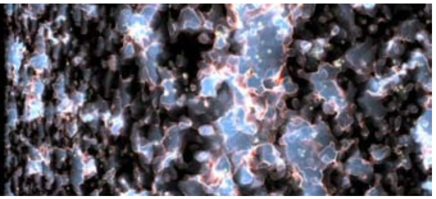
Computational Limitations

Harvard computers

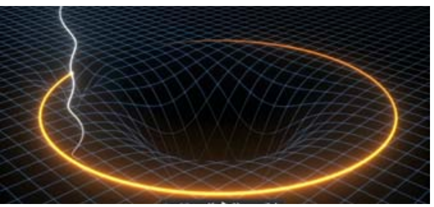
- Production of stellar plates and spectra ("data rate") was increasing enormously
- Examined and classified telescope output
- Forerunners of human mathematical computers



Key Science for the SKA
(a.k.a. m- and cm-λ astronomy)

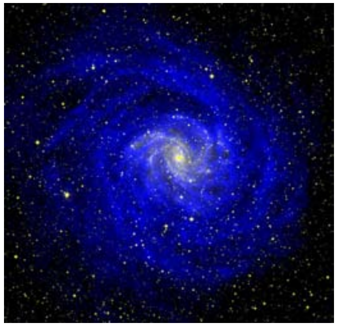


Emerging from the Dark Ages & the Epoch of Reionization

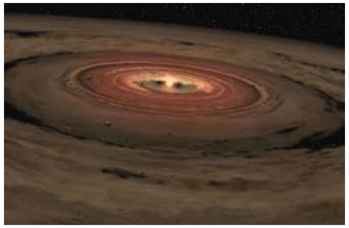


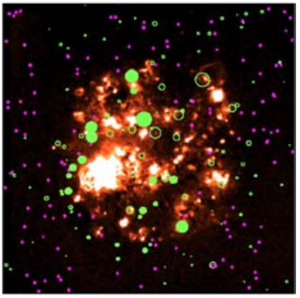
Strong-field Tests of Gravity with Pulsars and Black Holes

Galaxy Evolution, Cosmology, & Dark Energy



The Cradle of Life & Astrobiology



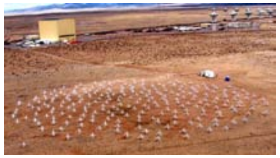


Origin & Evolution of Cosmic Magnetism

SKA Pathfinding







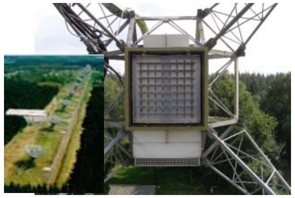

- SKA is ultimate goal, though long-term program
- Precursors and many pathfinders in existence or under construction

➢ Data challenges before SKA comes on-line

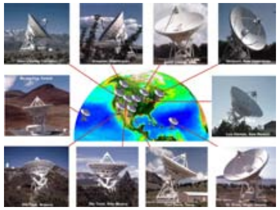
➢ Scalability could be an issue









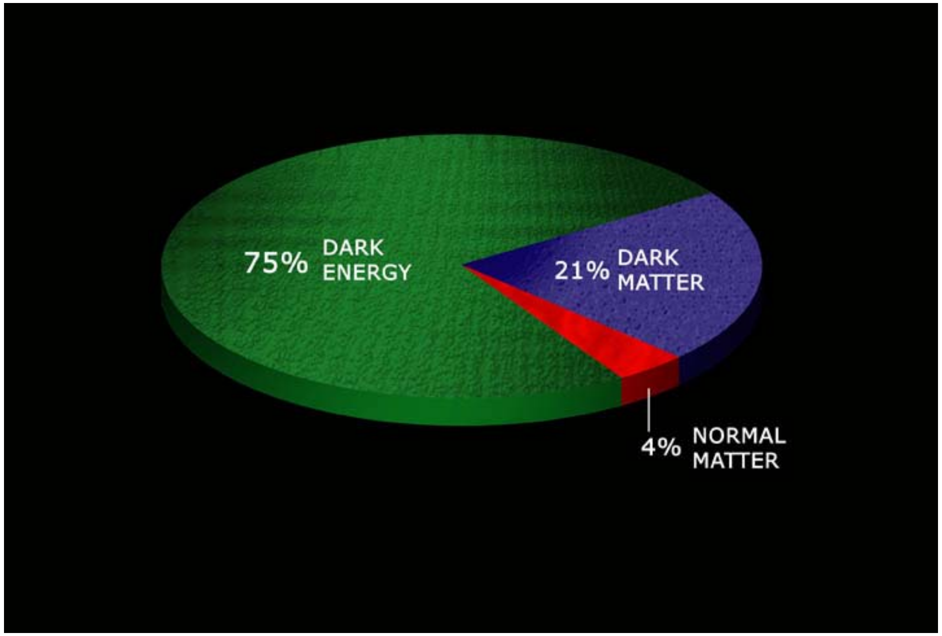




Case Study 1: Cosmology

Origin and Fate of the Universe

- Era of "precision cosmology"
  … or precision ignorance

- Need to sample a substantial volume of the Universe

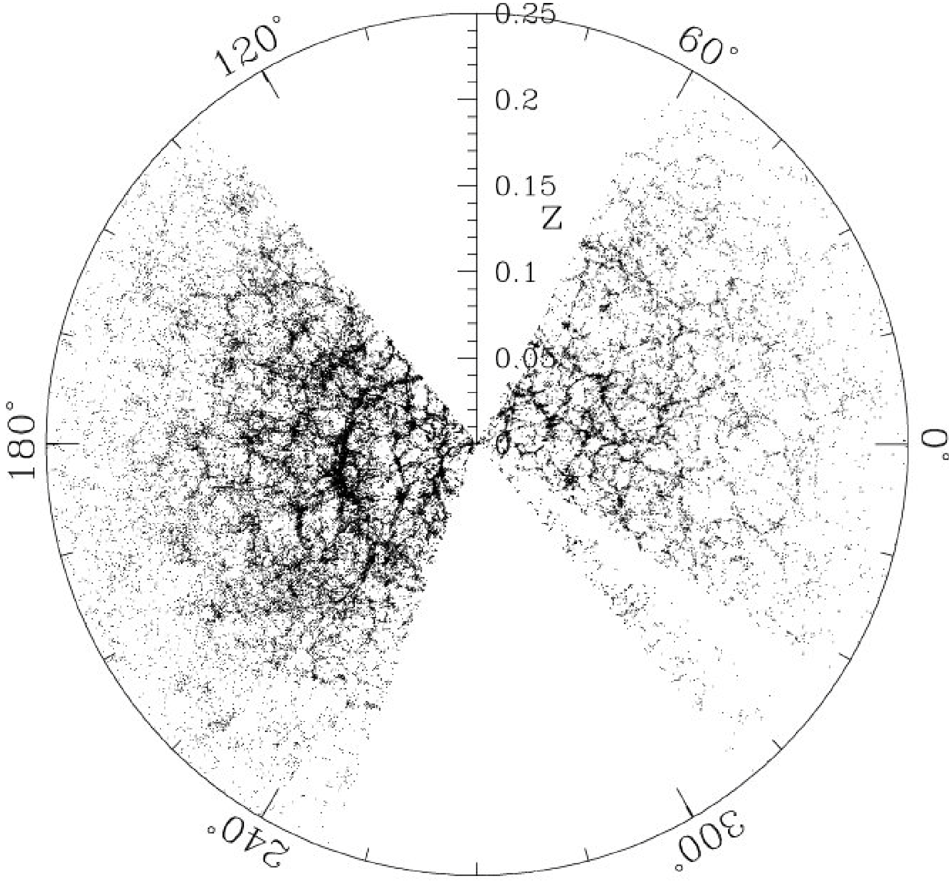


Composition of the Universe

Cosmology and Sky Surveys

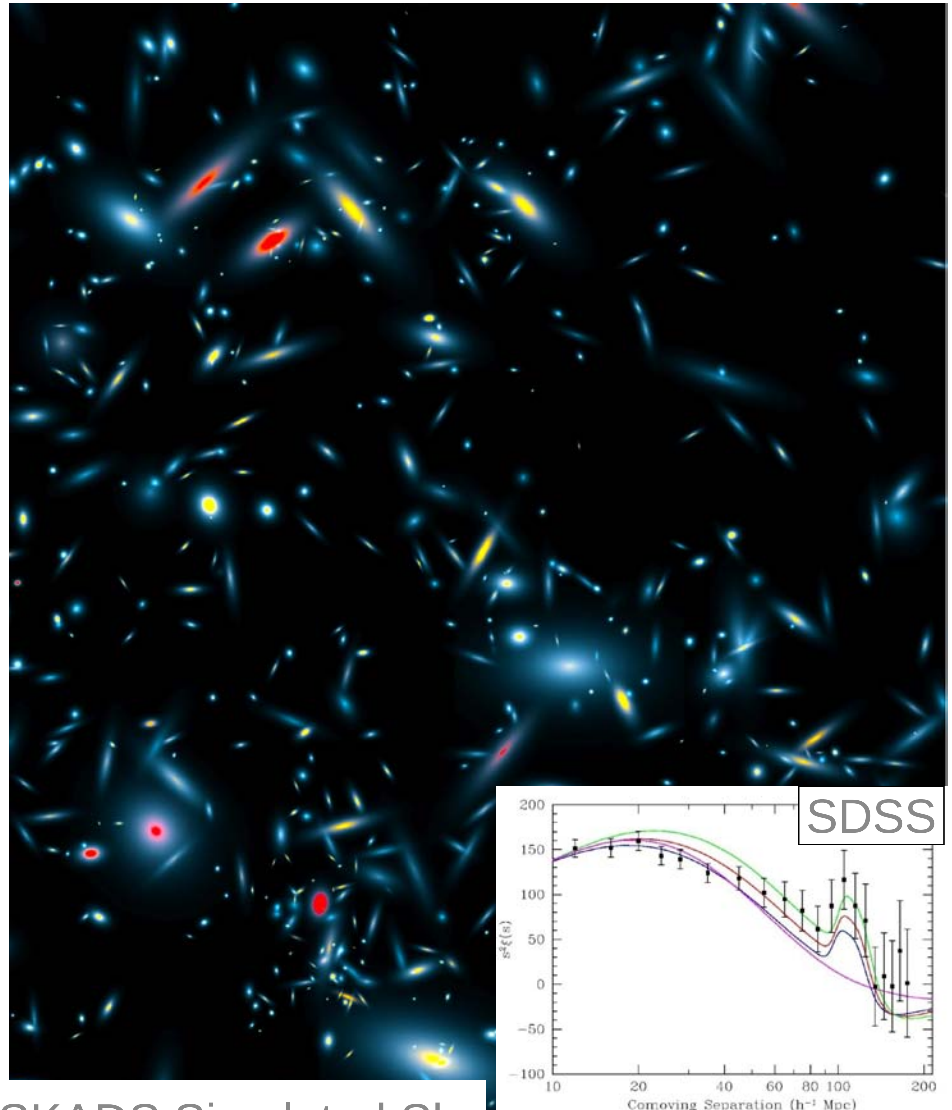
Volume ~ $D^2 \Delta D\ \Omega$

- D – distance; Ω – solid angle
- Surveying to larger D is difficult ➔ need larger telescopes
  - "square kilometre" of SKA
- Surveying larger sky areas Ω just requires more observing time



Sloan Digital Sky Survey volume

Cosmology and Sky Surveys



SKADS Simulated Sky

- Image the sky, locating galaxies

  Analysis of locations compared with cosmological models to constrain parameters

- Two broad classes of surveys
  - Continuum: e.g., NVSS, FIRST, *ASKAP/EMU, WSRT/APERTIF/WODAN*
  - Spectroscopic: SDSS, Arecibo ALFALFA, *ASKAP/WALLABY, SKA H I survey*

    Spectroscopic surveys locate in **3-D space!** very powerful

- Ultimate goal: spectroscopic survey of 1 billion galaxies

Imaging with Arrays

Fourier transform plane

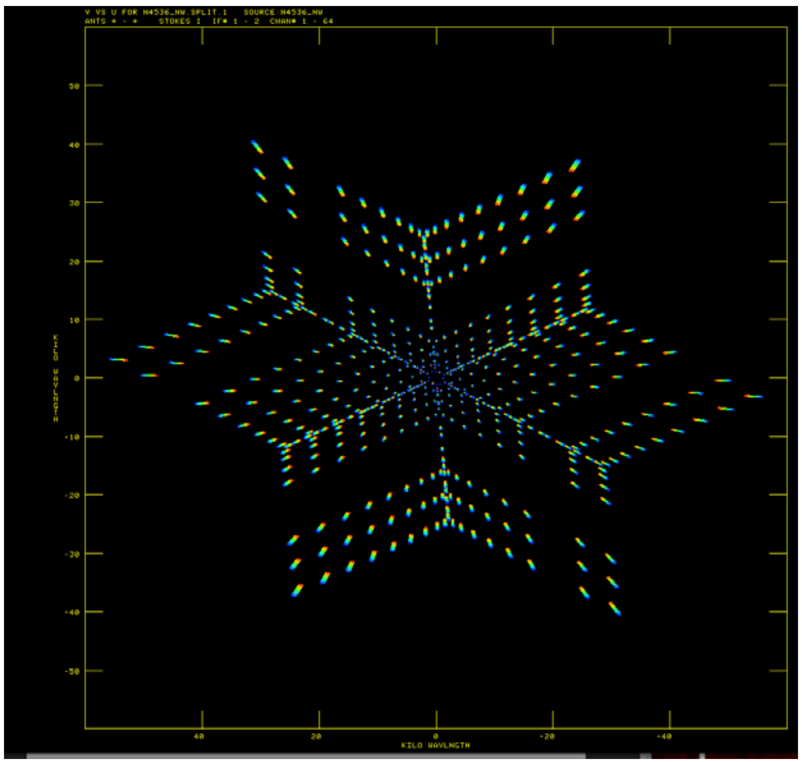
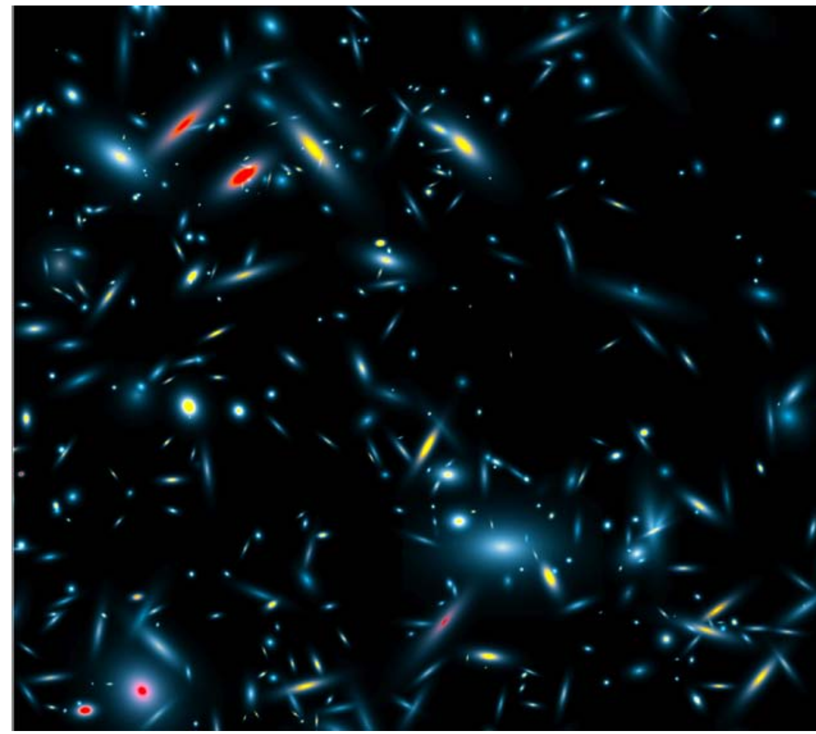


Image plane



Fourier Transform

←→

$$N_{data} \sim N_{antenna}^2 N_{frequency} N_{time}$$

Exploring the Universe with the world's largest radio telescope

Imaging Surveys

Requirements

- Many antennas in order to provide sensitivity and image quality
  large $N_{antenna}$
- Spectral resolution because of wide-field effects, line emission from galaxies, or both
  large $N_{frequency}$
- Long integrations in order to obtain adequate signal-to-noise ratio
  large N_{time}, e.g., 500 hr at 1 s sampling?

- $N_{data} \sim N_{antenna}^2 N_{frequency} N_{beams} N_{time}$

ASKAP	SKA Phase 1	SKA Phase 2
$N_{antenna} = 30$	$N_{anntena} \sim 250$	$N_{antenna} \sim 1000$
$N_{beams} = 30$	$N_{beams} = 1$	$N_{beams} = 1?$
$N_{frequency} \sim 16k$	$N_{frequency} \sim 16k?$	$N_{frequency} \sim 16k?$

Imaging Surveys II

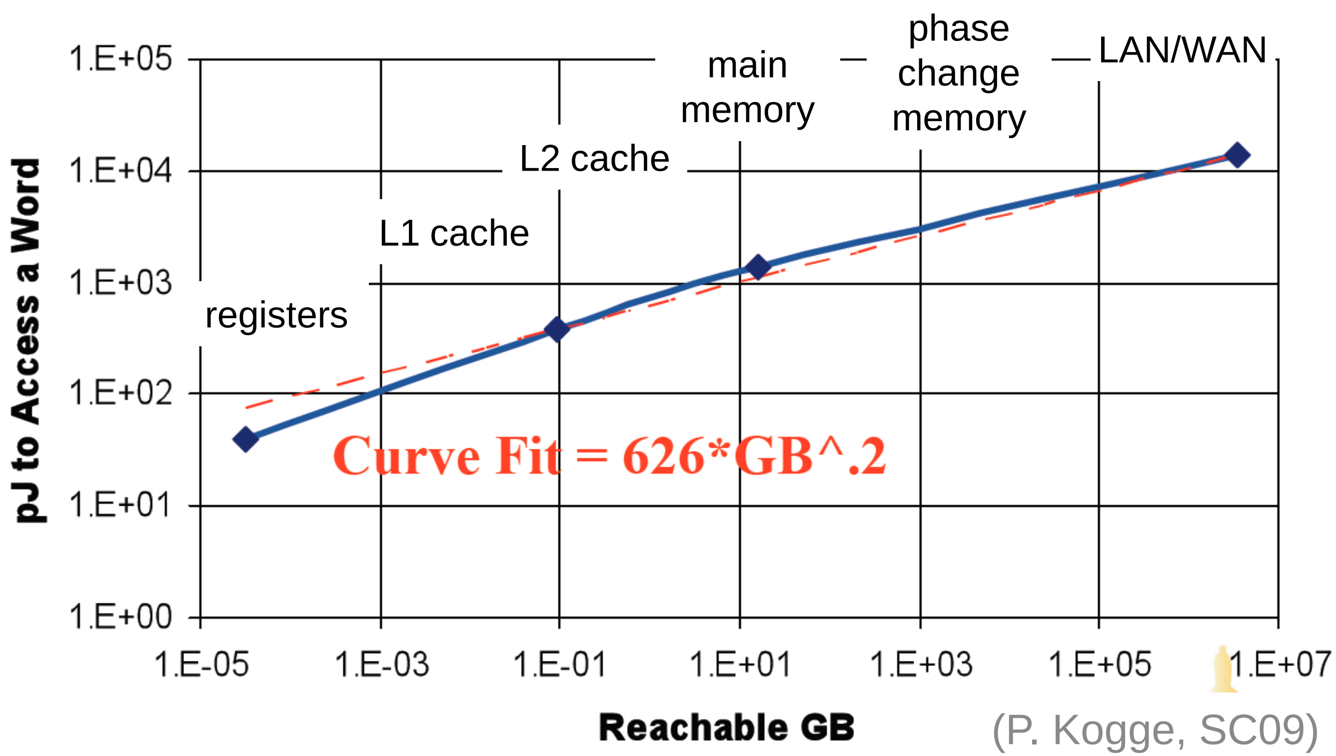
ASKAP	SKA Phase 1	SKA Phase 2
$N_{antenna} = 30$	$N_{anntena} \sim 250$	$N_{antenna} \sim 1000$
$N_{beams} = 30$	$N_{beams} = 1$	$N_{beams} = 1?$
$N_{frequency} \sim 16k$	$N_{frequency} \sim 16k?$	$N_{frequency} \sim 16k?$
$N_{time} \sim 1.8M$		
$N_{data} \sim 8 \times 10^{14}$	$N_{data} \sim 2 \times 10^{15}$	$N_{data} \sim 3 \times 10^{16}$
$\mathbf{N_{OPS} \sim 8 \times 10^{18}}$	$\mathbf{N_{OPS} \sim 2 \times 10^{19}}$	$\mathbf{N_{OPS} \sim 3 \times 10^{20}}$

- Imaging is more than “just” an FFT.
  - Gridding, deconvolution, wide-field corrections, self-calibration, …
- Community estimates are 10^4 to 10^5 ops per visibility datum(!)

Energy Consumption

- Really scary could be power consumption

  Early estimates for SKA power were 100 MW



(P. Kogge, SC09)

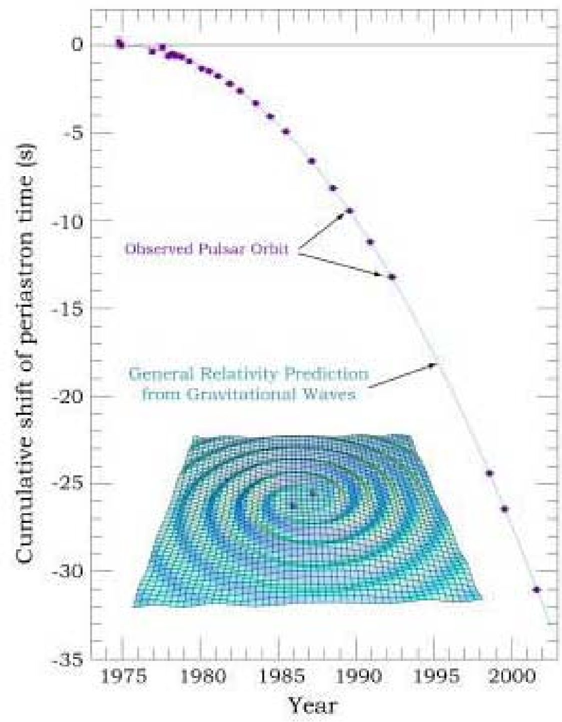
- Problem linked to moving data around chips
- See also D'Addario (SKA Memo #130)

Case Study 2: Fundamental Physics with Radio Pulsars

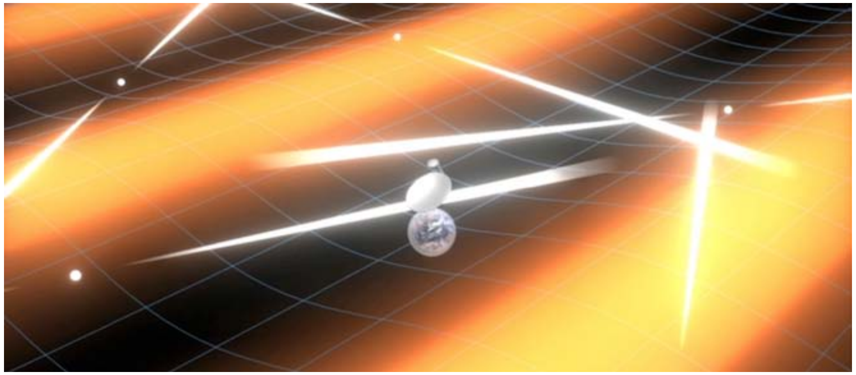
Arrival times of pulses from radio pulsars can be measured with phenomenal accuracy

- Better than 100 ns precision in best cases
- Enables high precision tests of fundamental physics
  - Theories of gravity, gravitational waves, nuclear equation of state
  - 1993 Nobel Prize in Physics
- Problem: Not all pulsars are equal!
- Good “timers” < 10% of total population
- Need to find **many** more!

➢ All-sky survey



← Ultra-relativistic binaries & gravitational wave studies



Pulsar Surveys I

Requirements

- Large bandwidths because pulsars are faint
- Long integration times because pulsars are faint
- Rapid time sampling in order to resolve pulse profile
- Narrow frequency channelization in order to mitigate interstellar scattering

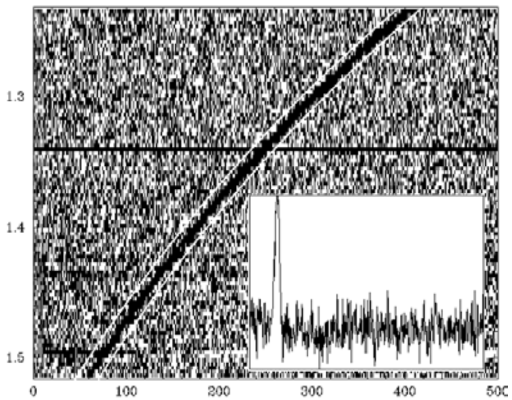
- For a “pixel” on the sky, accumulate data for a time Δt over a bandwidth $\Delta\nu$

    Suppose Δt = 20 min., $\Delta\nu$ = 800 MHz

- Time sampling δt with frequency channelization $\delta\nu$

    For GBT GUPPI, δt = 81.92 μs, $\delta\nu$ = 24 kHz

➢ 60 GB data sets per pixel …



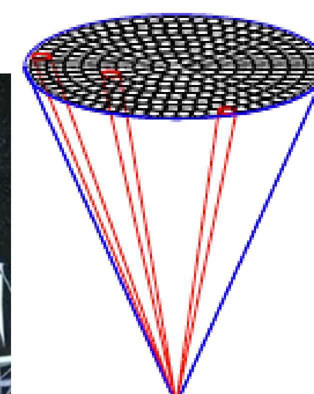
Pulsar Surveys II

For GBT

- At 800 MHz, "pixel" ~ 16' = 0.3°
- About 350 kpixels in the sky
- 20 PB data set

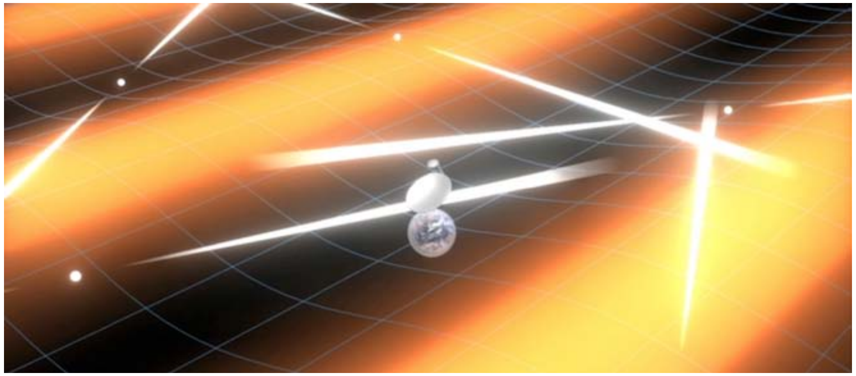
For SKA

- At 800 MHz, "pixel" = 1.2'
- About 76 Mpixels in the sky
- 4.6 EB data set



Case Study 3: Observing!

- Gravitational wave studies via a pulsar timing array
  - Precise measurements of arrival times of pulsars can track changes in pulsar-Earth distance



- Interferometric imaging
  - H I observations or continuum
  - Galaxy evolution through cosmic time, cosmology, magnetic field studies, …

Case 3a: Pulsar Timing

- Nyquist samping of a 1 GHz bandwidth = 2 Gsamples/s
- 4 polarizations
- 1-byte representation of data?
- 15 minutes per pulsar
- Weekly observations of 100 pulsars

  (for 5 to 10 years)

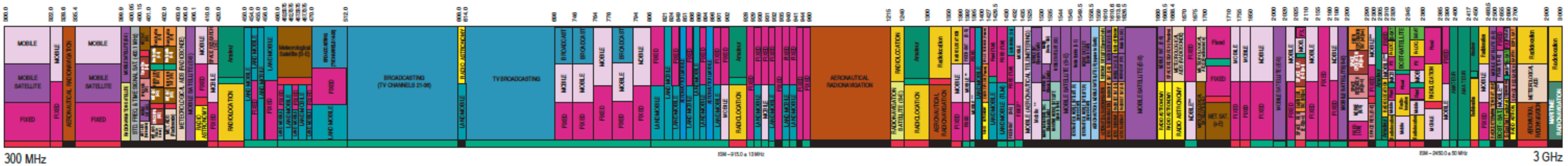
- **= 3 PB/month**

Case 3b: Imaging

- Nyquist samping of a 1 GHz bandwidth = 2 Gsamples/s
- 4 polarizations
- 1000-element array (499,500 visibilities)
- 1-byte representation of data?

- **= 4 PB/s**

Case 4: Interference Mitigation



- Most of the radio frequency (RF) spectrum is not reserved for use by radio astronomy
  - In fact, very little is! ☹
- Other passive users are fine
- Active users can be troublesome!
  - GPS, microwave ovens, cell phones, car ignitions, electric fences, …

Interference Mitigation II

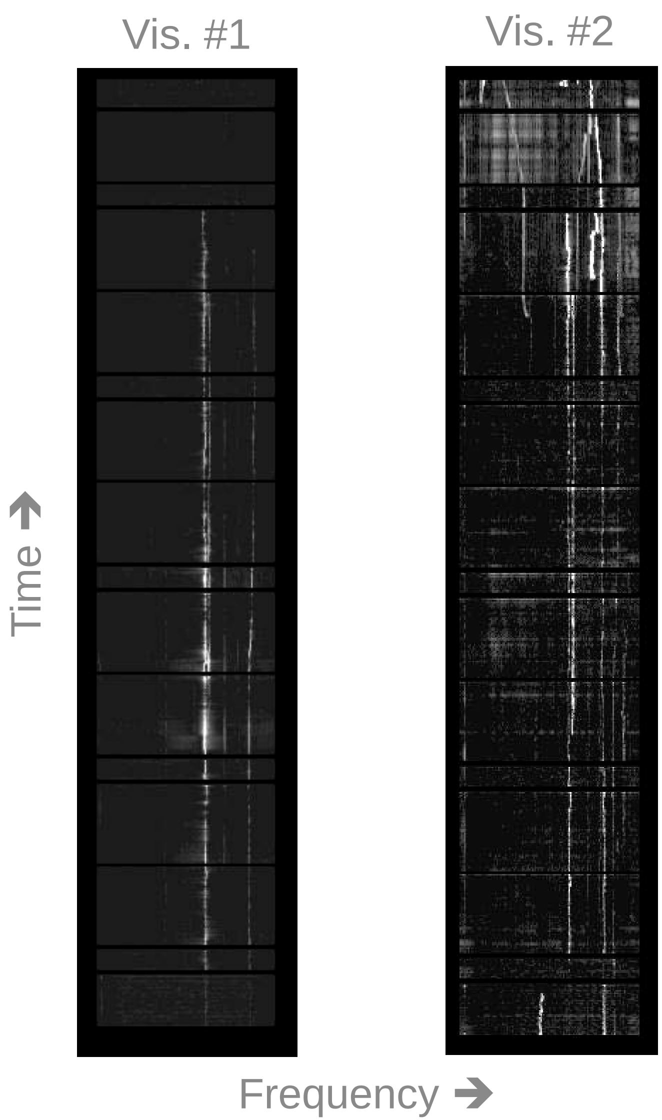
- Radio flux density measured in Janskys

  1 Jy = 10^{-26} W/m^2/Hz

- 10 μJy = EVLA, GBT, ASKAP, MeerKAT, … sensitivity
- 10 nJy = SKA aim

- 100 Jy ~ cell phone on Moon

Data Visualization



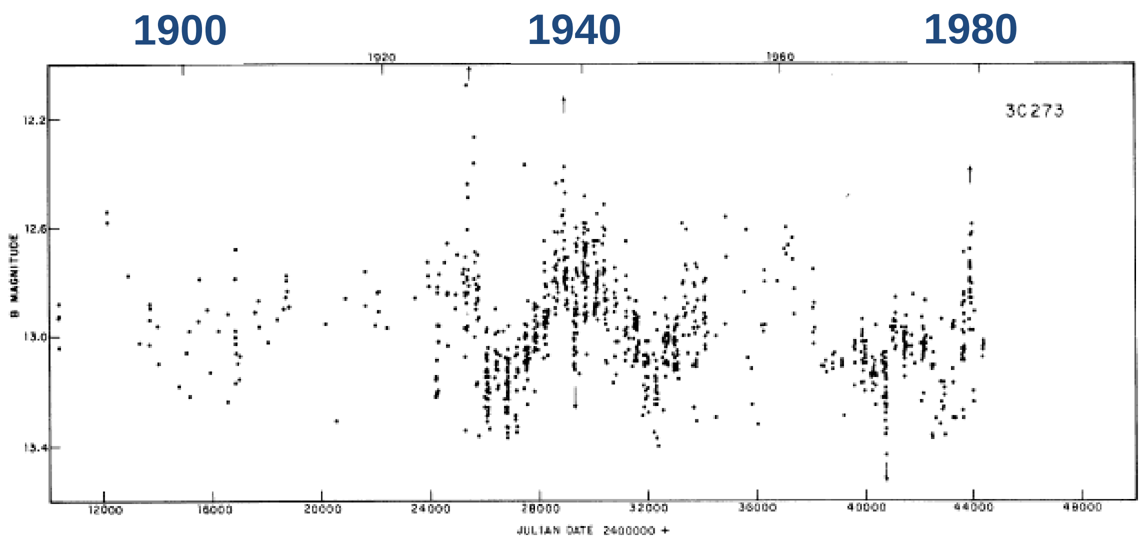
- Data acquired from an array are at least 4-D
  - Visibility (= $\text{antenna}_i \times \text{antenna}_j$)
  - Frequency
  - Time
  - Polarization
  - (Beams? for a multi-beam system)
- Best results still obtained by hand

Case 5: Astronomy for the Future

- 100+ year light curves
- BL Lac
  - Variable star #90 in constellation Lacertae?
  - Active galactic nucleus?

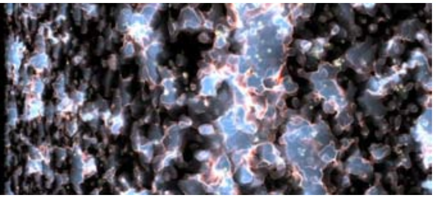
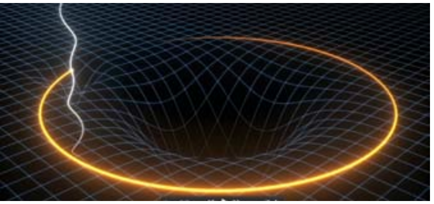
➢ What will our "academic grand-children" want to know?

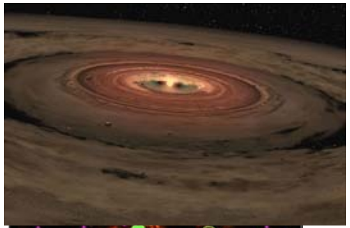
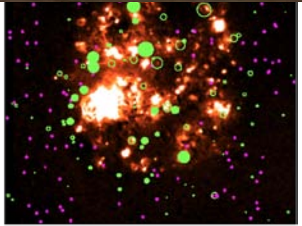
➢ What will they be able to know?



3C 273 (Angione & Smith 1985)

Summary – Data-Intensive Astronomy

- Exciting science!
- Leads to exciting data challenges
  - Data volume (Exabyte)
  - Processing requirements (Exo-flop)
  - Data rates (PB/s)
  - Data visualization (HMI)
  - Long time scales (100 years?)
- Answers on Thursday

Pulsar Surveys III

- Processing load