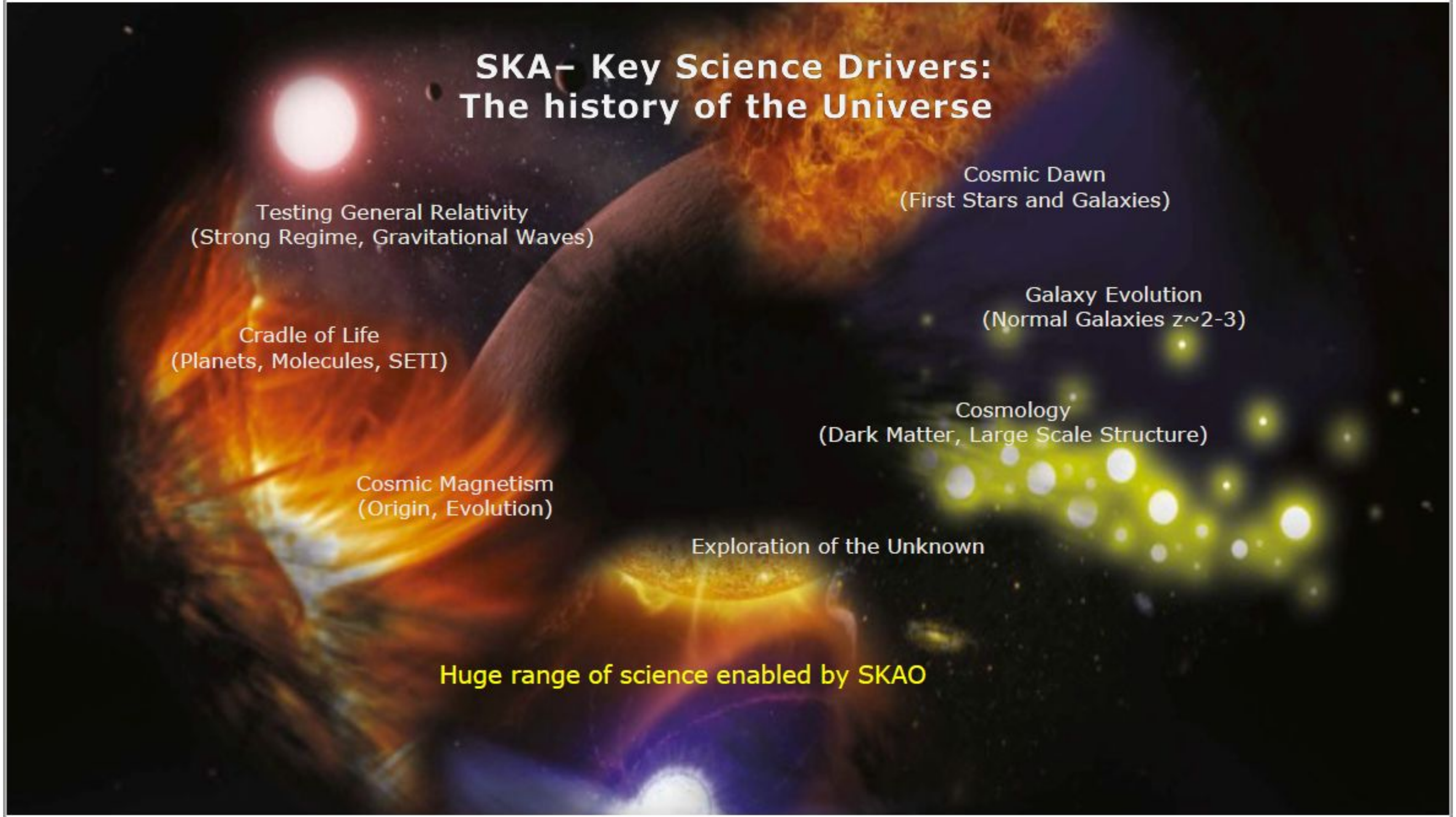
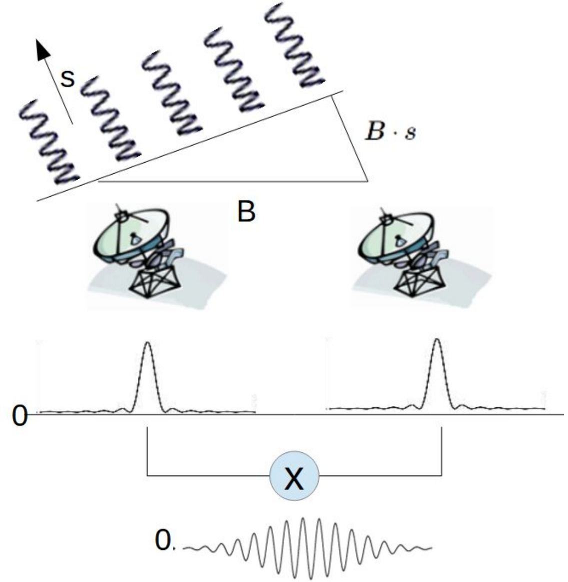
# Towards the SKA: The origin of big data

**Marcella Massardi**
(INAF- IRA - Italian ARC - SISSA)

SRC training event - January 2022


SKA– Key Science Drivers:
The history of the Universe
Cosmic Dawn
(First Stars and Galaxies)
Testing General Relativity
(Strong Regime, Gravitational Waves)
Galaxy Evolution
(Normal Galaxies z~2-3)
Cradle of Life
(Planets, Molecules, SETI)
Cosmology
(Dark Matter, Large Scale Structure)
Cosmic Magnetism
(Origin, Evolution)
Exploration of the Unknown
Huge range of science enabled by SKAO

# Where does the big data come from?




A minimum radio-mm interferometer is a couple of antennas that collects amplitude and phase of an electromagnetic signal on frequencies<1THz

The signals collected from each antennas are correlated to amplify signal and reduce noise. **The correlation product is a visibility**

Source shape

$$R(\boldsymbol{B}) = \iint_{\Omega} A(\boldsymbol{s}) I_{\nu}(\boldsymbol{s}) \exp\left[\mathrm{i}\, 2\pi\nu \left(\frac{1}{c}\boldsymbol{B} \cdot \boldsymbol{s}\right)\right] \mathrm{d}\Omega\, \mathrm{d}\nu$$

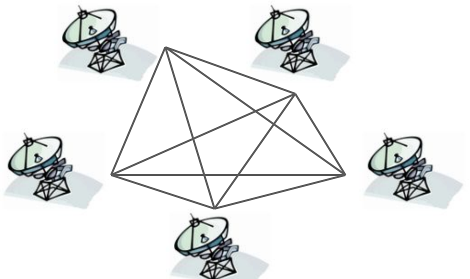
Frequency

AMPLITUDE: Source brightness modulated by the telescope response function

PHASE: Projection of the antenna distances in the source direction

Hence the interferometer collects

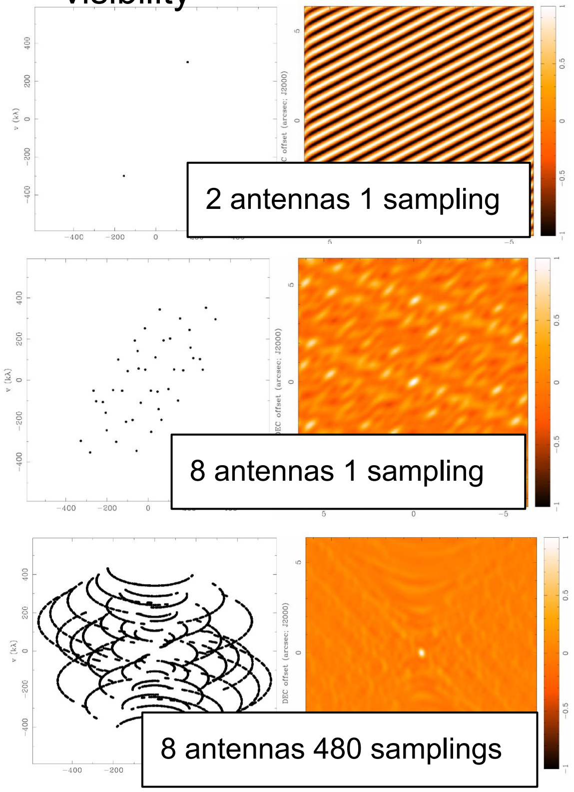
- 2 values (amp and phase)
- x each couple of antennas
  - (e.g. N antennas means N(N-1)/2 independent couples)
- x each polarization
  - (up to 2 perpendicular directions)
- x each frequency channel
- x each timestamp (typically 10sec)



# Where does the big data come from?

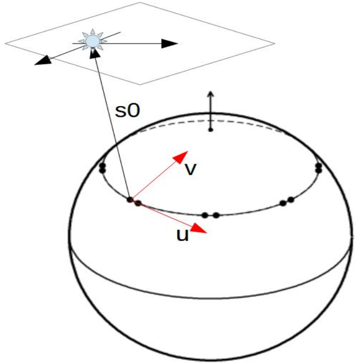
sampled visibility

Image




$$R(\boldsymbol{B}) = \iint_{\Omega} A(\boldsymbol{s}) I_{\nu}(\boldsymbol{s}) \exp\left[ \mathrm{i}\, 2\pi\nu \left( \frac{1}{c} \boldsymbol{B} \cdot \boldsymbol{s} \right) \right] \mathrm{d}\Omega\, \mathrm{d}\nu$$

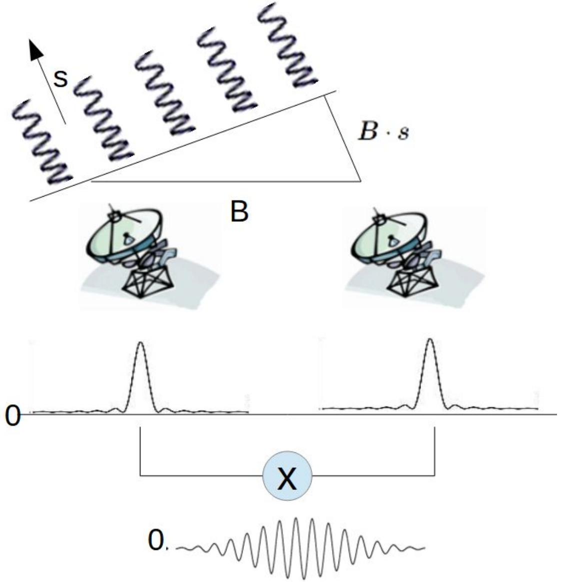
A visibility is the Fourier transform of the brightness distribution of a signal.
The Interferometer samples amplitude and phase in the Fourier domain and an IFT is necessary to reproduce the image.




-> The larger the number of antenna couples
the better is the coverage of the visibility domain in shorter observing time
and the better is the possibility to reconstruct the image with a
non linear process of image reconstruction

-> The longer the distance between antennas in a couple,
the higher is the resolution of the image
(i.e. smaller details, larger number of pixels, larger size of images)

# Where does the big data come from?




**Telescope sensitivity**

Brightness temperature corresponding to all the signals collected including source, atmosphere and instrument

Boltzmann k

$$\sigma_S \approx \frac{2\,k\,T_{sys}}{A_{eff}\,\sqrt{n(n-1)\times\Delta\nu\times\eta_{pol}\times t_{int}}}\ [\mathrm{Jy}]$$

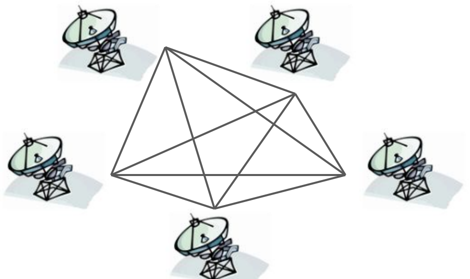
Time on source

# of polarizations

Bandwidth

Effective collecting Area= dish_area x efficiency

n = # of antennas
n(n-1)/2 = # of baselines



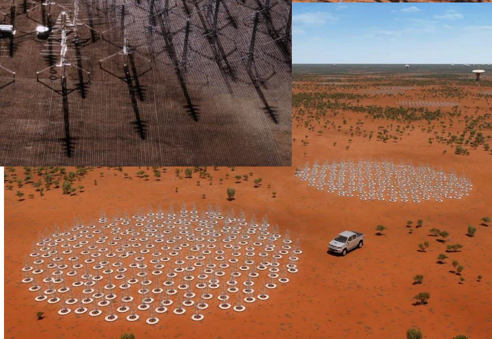
The larger the number of antenna couples, polarizations, bandwidth, time
the lower is the noise we get, higher the signal to noise ratio,
the better is the possibility to detect fainter signals

# Square Kilometer Array

in South Africa
dishes
0.35<frequencies <15 GHz
baselines up to 150 km

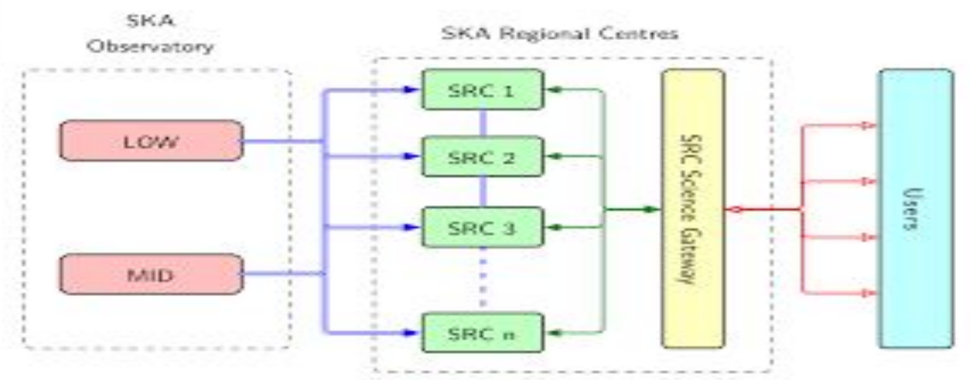
in Western Australia 197
512x256 linear elements
frequencies 50-350 MHz
baselines up to 65km

Phase 1 completed in 2030



SKA Phase 2: 2500 dishes across Africa;
1,000,000 antennas across Australia

# SKA Organization

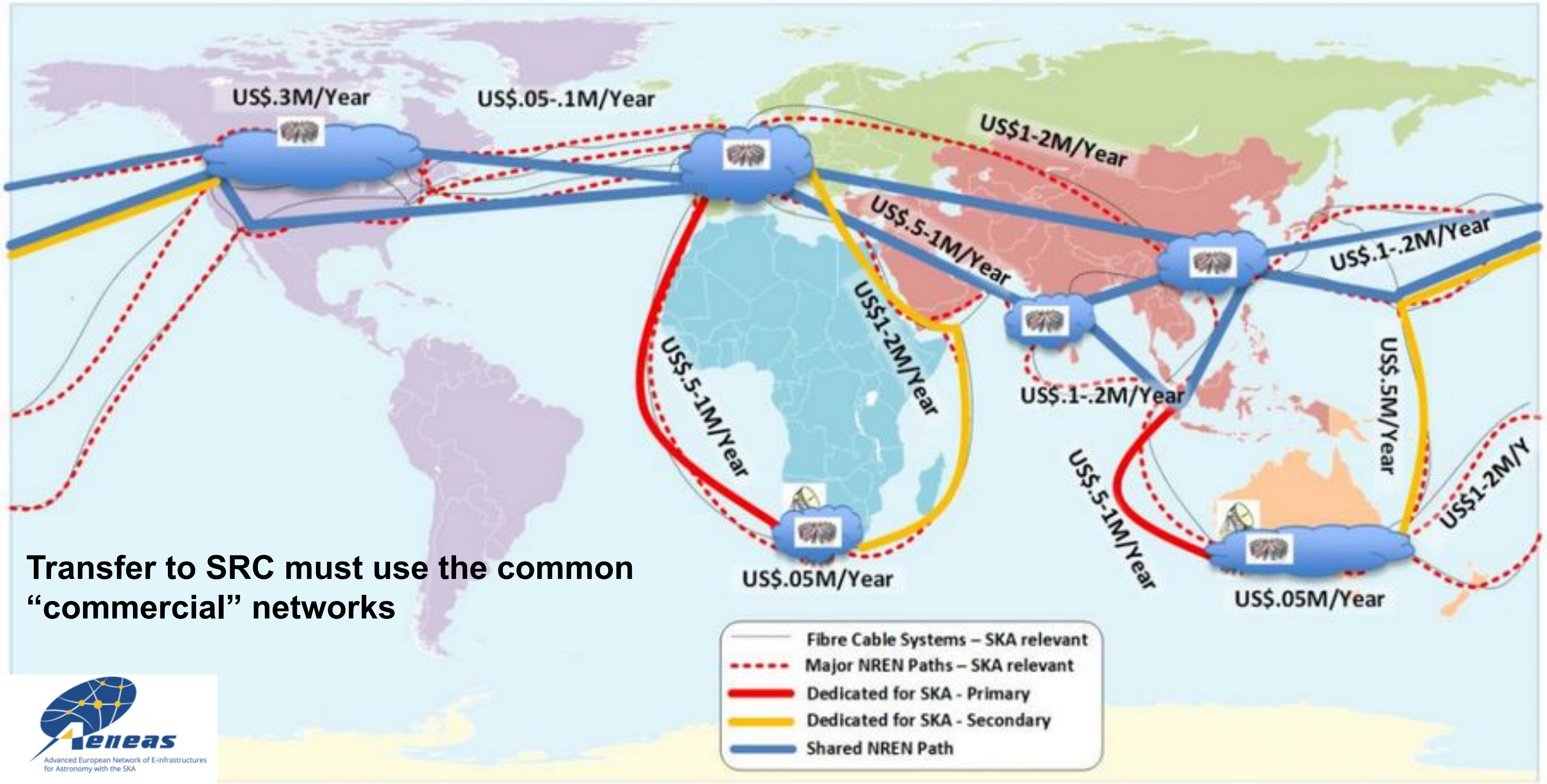

The SKA organization manages the telescopes construction and telescope data distribution.

The members will organize with SKA Regional Centres to receive, analyze and store the advanced products.

**Community is divided into Science Working Groups to identify instrumental requirements that might enhance the scientific return.**

- Science Working Groups & Focus Groups
  - Cosmology
  - Cradle of Life
  - Epoch of Reionization
  - Extragalactic Continuum (galaxies/AGN, galaxy clusters)
  - Extragalactic Spectral Line
  - Gravitational Waves
  - HI galaxy science
  - Magnetism
  - Our Galaxy
  - Pulsars
  - Solar, Heliospheric & Ionospheric Physics
  - Transients
  - High Energy Cosmic Particles (Focus Group)
  - VLBI

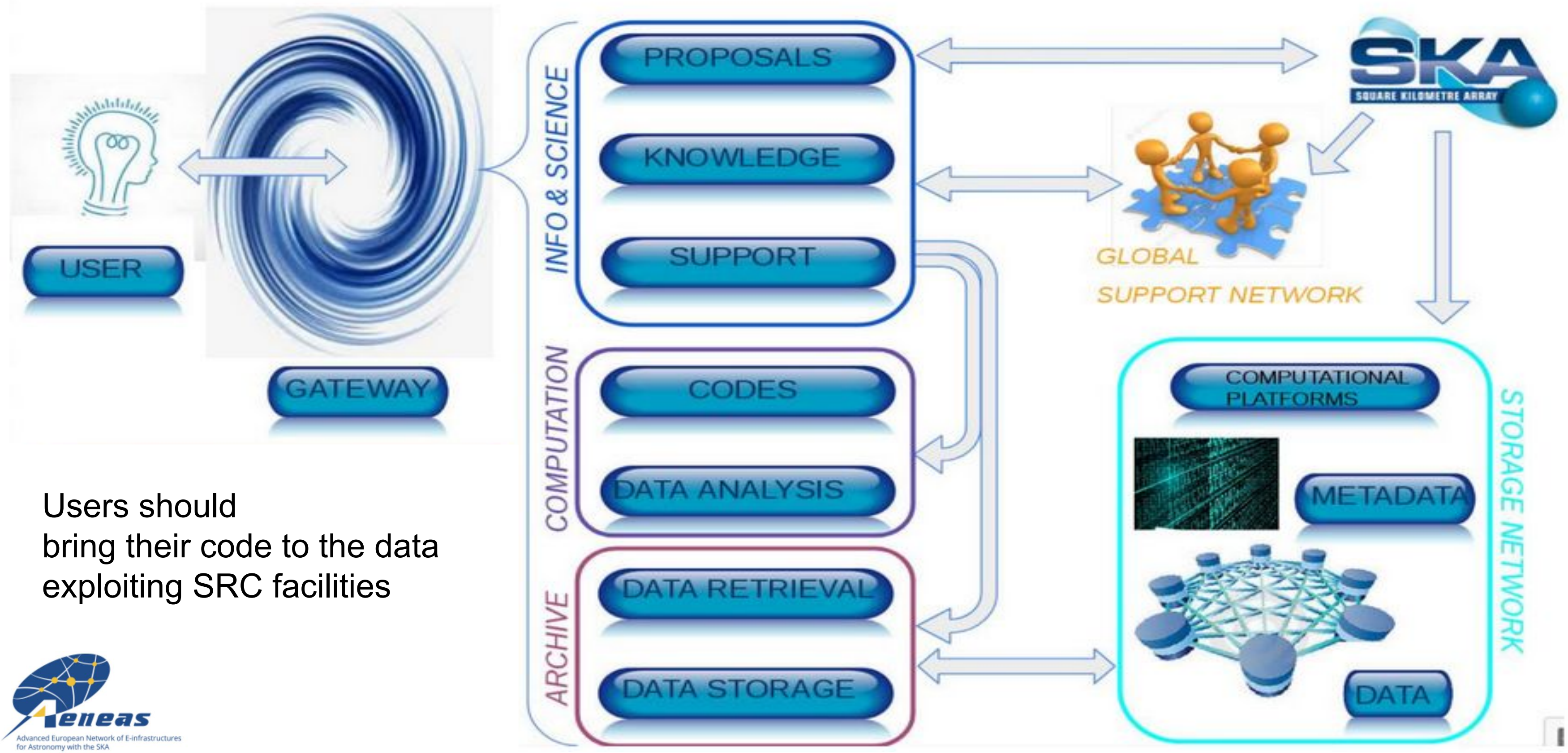
# SKA Data flow




**Transfer to SRC must use the common "commercial" networks**

Advanced European Network of E-infrastructures for Astronomy with the SKA

# SKA Regional Centres




Users should
bring their code to the data
exploiting SRC facilities

Advanced European Network of E-infrastructures for Astronomy with the SKA

# Square Kilometer Array

| **Nominal Frequency** | **110 MHz** | **300 MHz** | **770 MHz** | **1.4 GHz** | **6.7 GHz** | **12.5 GHz** |
|---|---|---|---|---|---|---|
| **Range [GHz]** | 0.05-0.35 | 0.05-0.35 | 0.35-1.05 | 0.95-1.76 | 4.6-8.5 | 8.3-15.3 |
| Telescope | Low | Low | Mid | Mid | Mid | Mid |
| FoV [arcmin] | 327 | 120 | 109 | 60 | 12.5 | 6.7 |
| Max. Resolution [arcsec] | 11 | 4 | 0.7 | 0.4 | 0.08 | 0.04 |
| Max. Bandwdith [MHz] | 300 | 300 | 700 | 810 | 3900 | 2 x 2500 |
| Cont. rms, 1 hr [μJy/beam][a] | 26 | 14 | 4.4 | 2 | 1.3 | 1.2 |
| Line rms, 1 hr [μJy/beam][b] | 1850 | 800 | 300 | 140 | 90 | 85 |
| Resolution Range for Cont. and Line rms [arcsec][c] | 12–600 | 6–300 | 1–145 | 0.6–78 | 0.13–17 | 0.07–9 |
| Channel width (uniform resolution across max. bandwidth) [kHz] | 5.4 | 5.4 | 13.4 | 13.4 | 80.6 | 80.6 |

# SKA Data rate:raw data

$$R(\boldsymbol{B}) = \iint_{\Omega} A(\boldsymbol{s}) I_\nu(\boldsymbol{s}) \exp\left[\mathrm{i}\, 2\pi\nu \left(\frac{1}{c}\boldsymbol{B}\cdot\boldsymbol{s}\right)\right] \mathrm{d}\Omega\, \mathrm{d}\nu$$

Hence the interferometer collects

2 values

(amp and phase -> 8By per integer number)

x each couple of antennas

(e.g. N antennas means N*(N-1) independent couples)

x each polarization product

(up to 2 perpendicular directions)

x each frequency channel

(e.g. 810MHz/13.4kHz~ 2)

x each timestamp

(typically about 10s)

for SKA MID in a **6hr observations** =

2*8By*(200*199)couples*2pol*$2^{16}$channels*360timestamps/h*6h=**180PB**

time to transfer at 100Gbit/s=>**4000h** !!!!!

# SKA Data rate: images


Image resolution:

consider at least 3-7 pixels in a synthesized beam (e.g. 0.4 arcsec)

Image size:

consider at least 2 times the FoV (e.g. 60 arcmin)

**npixels per beam = im_size/im_resolution**~ $2\,\lambda/D \cdot 1/(5 \cdot \lambda/B_{max}) \sim 10 \cdot B_{max}/D$

**for a single image at 1.4GHz**

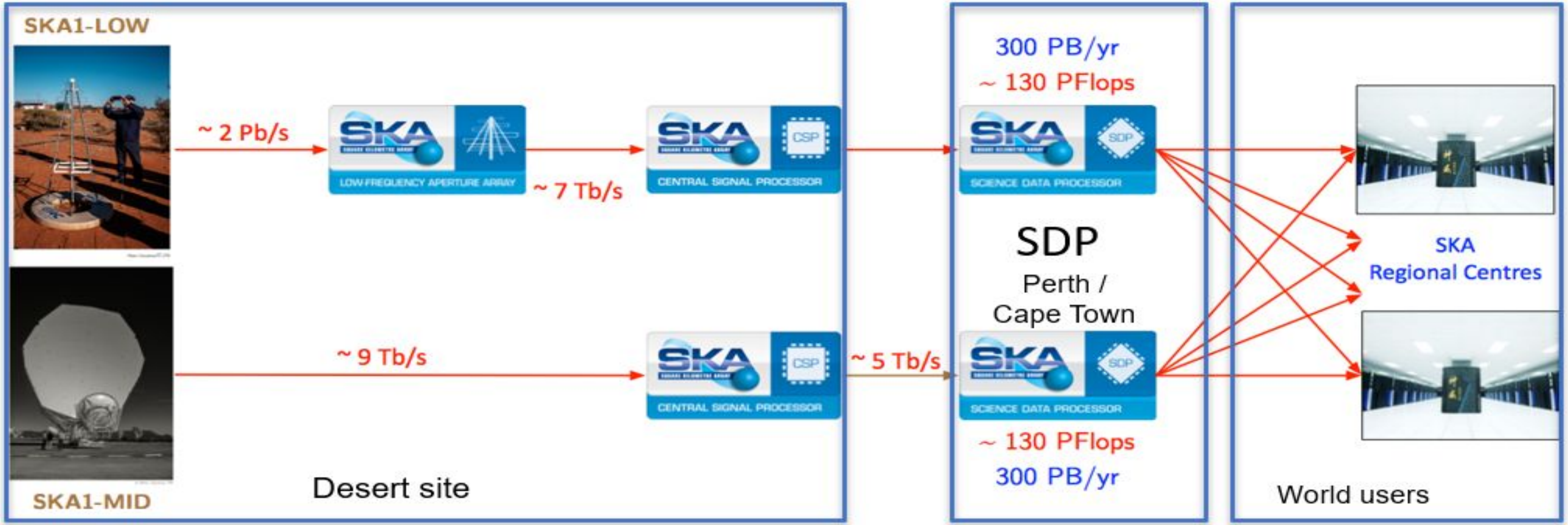
=> image pixels per side= 120arcmin/0.08 arcsec

=> image size= 90000x90000 pixels*8B=

**65GB/channel** (up to $2^{16}$ channels mean 4PB)

# Square Kilometer Array

# SKA Data flow

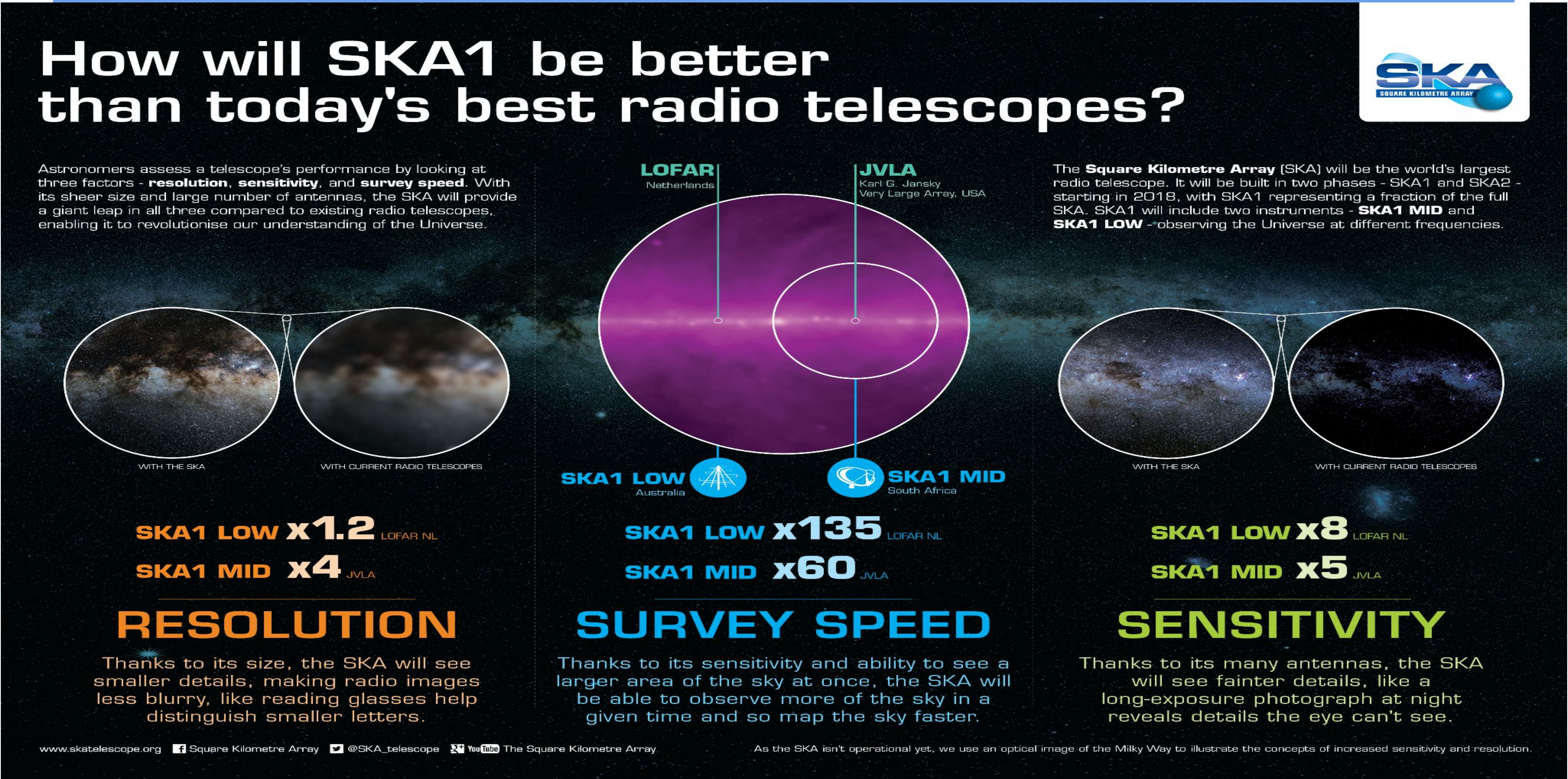



- But, have a series of buffers
- *Raw voltage data can be stored for about 2 minutes*

- *Raw visibility data can be stored for about 2 weeks*

- *Final data products will be stored forever*

# Square Kilometer Array



# How to get ready to SKA?

**From a global perspective:**

-Building the system on science/community based requirements

-Identify the potential problems for the future (e.g. data transfer, satellite constellations, atmospheric communication)

**From the astronomer perspective:**

-Test on pathfinders/precursors/simulations (e.i. data challenges)

-Improve on big data handling skills

-Enter a Science Working Group