



MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS



Bancos de Dados Geográficos

#10 – Sistemas de Bancos de Dados Geoespaciais (Dados Matriciais)

Dr. Gilberto Ribeiro de Queiroz <gribeiro@dpi.inpe.br>

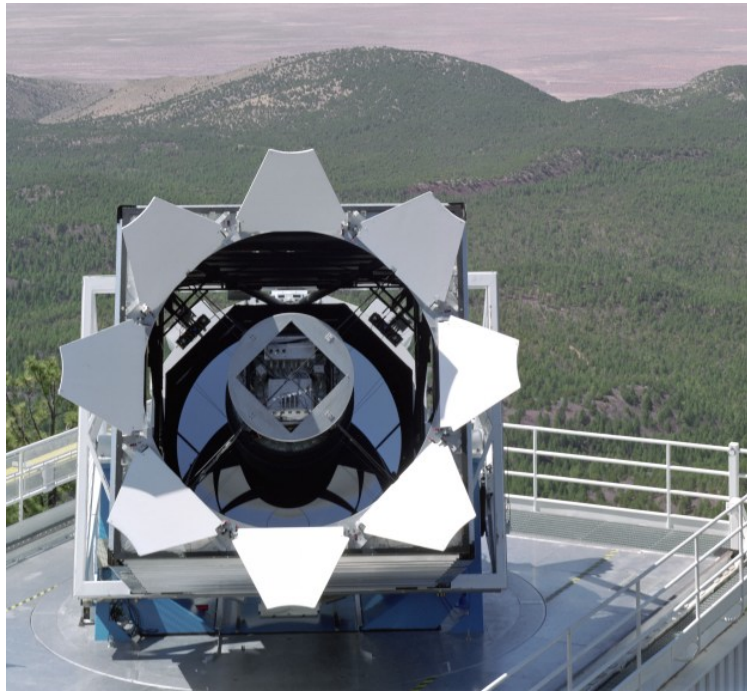
Dr. Eymar Lopes <eymar@dpi.inpe.br>

Introdução às Imagens de Sensoriamento Remoto

Os dados utilizados em diversas áreas da
Ciência encontram-se na forma de
Arrays

Arrays = Matrizes

Imagens de Telescópios Astronômicos



[The SDSS survey telescope](#)
(Sacramento Mountains)

Fonte: [The Sloan Digital Sky Survey](#)

[Sloan Digital Sky Survey / SkyServer](#)

- Em operação desde 2000
- SDSS I e II (2000-2008):
 - 10 TiB raw data, 120 TiB processados
- [SDSS III](#)

Fonte: Takar (2013)

[Messier 51, The Whirlpool Galaxy](#)



Biologia: Genoma, Máquinas de Sequenciamento

```
GATCAGAGAAATTCCAGCATATGACATCCACG
CGCTAGCCGGTATATGAAATGAGAGGATCATC
ACACTATGTGATGACATACTAGACC GG TGATG
GGGATATCAGGAATTCCAGCATATGACATCCA
CGCGCTAGCCGGTATATGAAGGATGAGAGGGA
GCCACCACTATGTGATGACATACTAGACC GG T
ACGATGGATTACAGGAATTCCAGCATATGACA
GAGGCCACCGCTAGCCGCTATATGAAATGAG
AGAGGGACACCACTATGTGATGACATACTAGA
CCC GG TGATGGATTACAGGAATCCAGCATA
TGACATACACCGCTAGCCGAGTATATGAGAG
ACATGAGAGGGACACCACTATGTGATGACATA
CCCTAGACC GG TGATGGATTACAGGAATTCCC
GCATATGACACCCACCGCTAGCACGTATAAG
CATTGAAATGAGAGAGGAATCCACTATGTGAT
GACATACTAGACC GTTGTGATGGATTACAGG
AATCCAGCATATGACATCCACATCCCTAGCTC
CAGGTATATGAAATGAGAGGGACACCACTATG
```

INDEX

AAA OFFSETS: 9, 49, 257, 467, 571

ATC OFFSETS: 2, 60, 104, 127, 319, 480, 551

CGG OFFSETS: 40, 124, 141, 194, 300, 404

DNA -> blueprint for all living organisms

Genomics: Source of big data

Large dense matrix of floating point numbers

Analytics: Regression, Covariance, Biclustering, SVD, Statistics

Fonte: Taft et al. (2013)

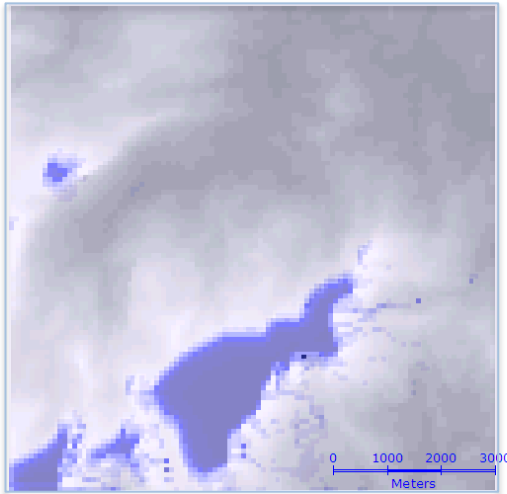
Some projects:

- ✓ [The Human Genome Project](#) (HGP)
- ✓ [1000 Genomes](#)
- ✓ [Stanford Microarray database](#)

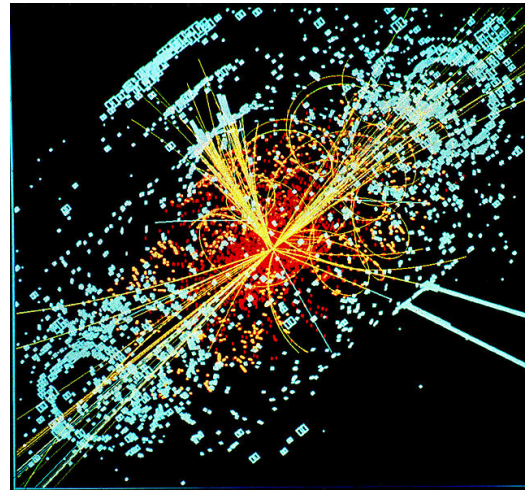
Fonte: [Schatz and Langmead \(2013\)](#)

Simulação Computacional

Fonte: (Carneiro, 2006)

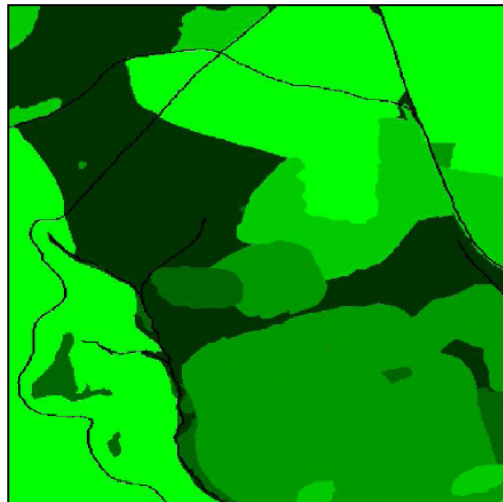


Fonte: [Wikipedia](#)

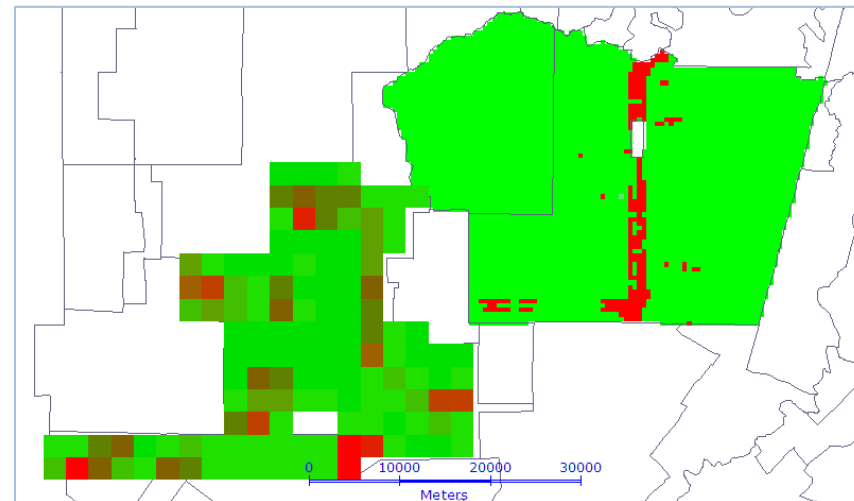


Simulated Large Hadron Collider CMS particle detector data depicting a Higgs boson produced by colliding protons decaying into hadron jets and electrons

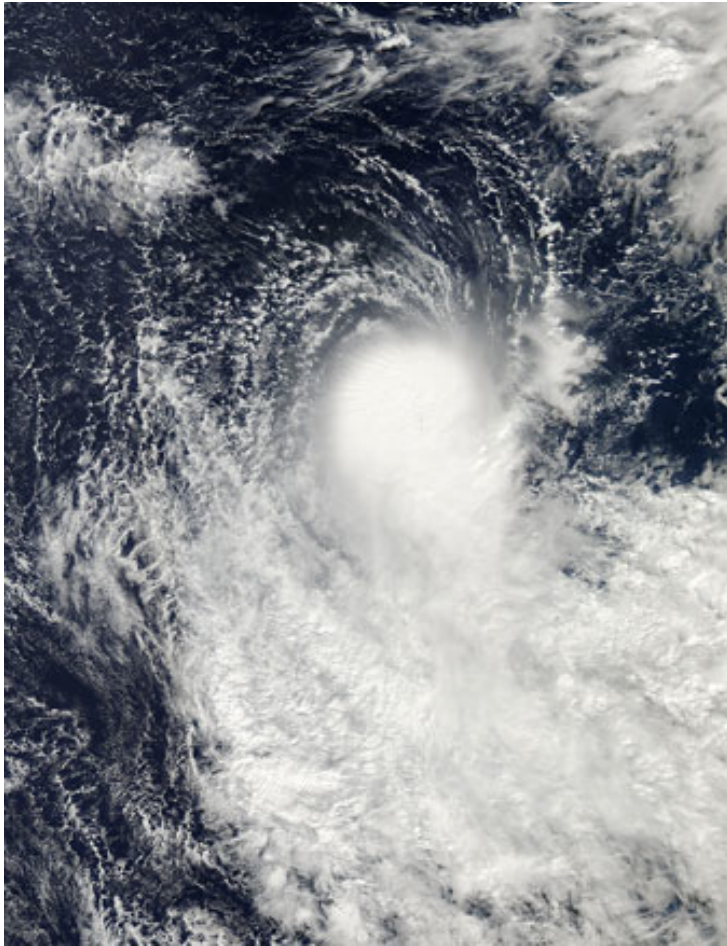
Fonte: (Almeida et al, 2008)



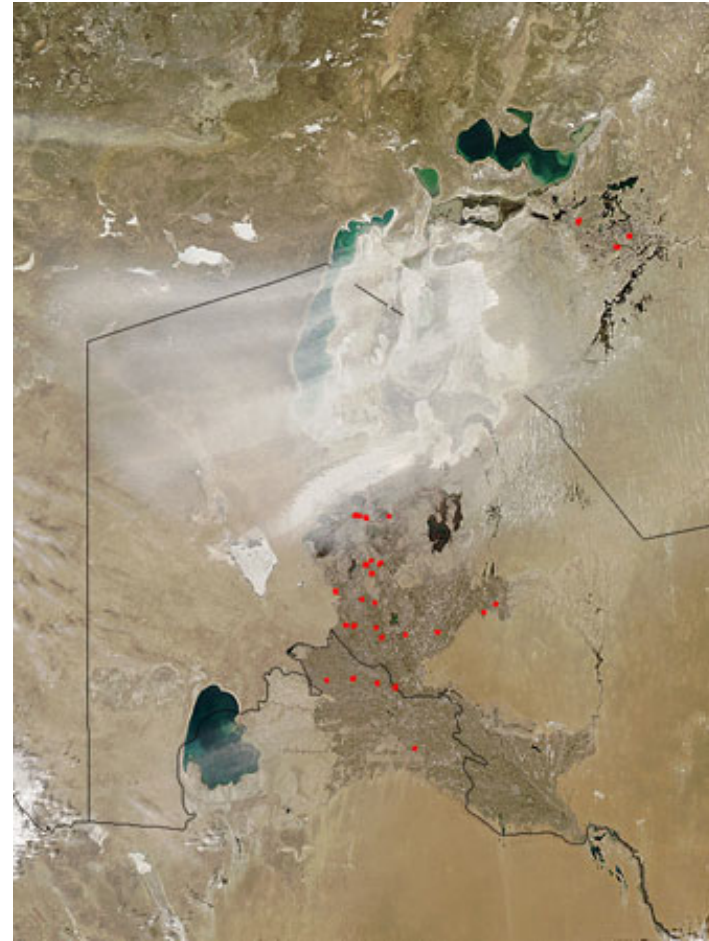
Fonte: (Carneiro, 2006)



Dados Geoespaciais: imagens de satélite



Tropical Cyclone Jack (24S) in the South Indian Ocean
Fonte: <http://modis.gsfc.nasa.gov> (Abril, 2014)



Spreading dust storm over the Aral Sea
Fonte: <http://modis.gsfc.nasa.gov> (Abril, 2014)

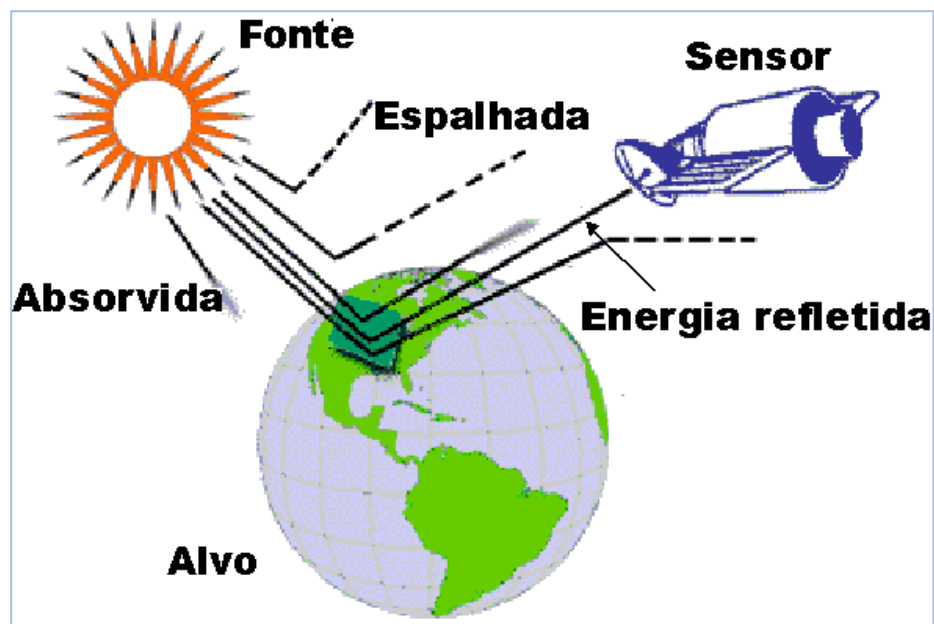
Arrays: A.K.A.

- Raster data
- Regularly gridded data
- Multi-Dimensional Discrete Data (MDD)
- Imagens
- Dados Matriciais

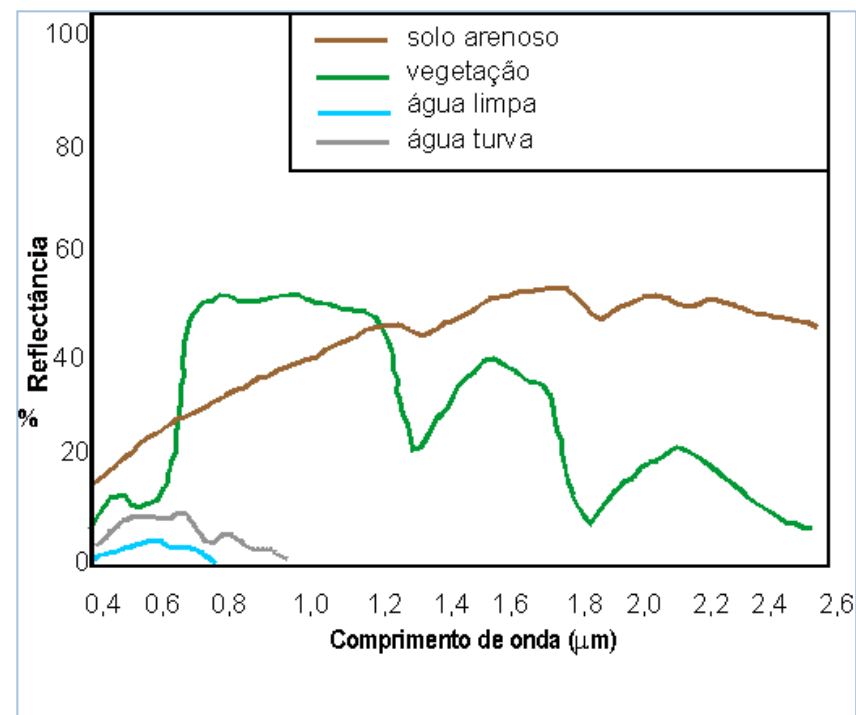
Como as imagens são organizadas?

- O que é uma imagem de sensoriamento remoto?
- Como os dados de sensoriamento remoto são representados?
- Quais os principais formatos de imagem?
- Os 2V's associados a dados de sensoriamento remoto.

Coleta de Dados: Sensoriamento Remoto (Imagens Ópticas)



Os sensores medem a quantidade de energia refletida ou emitida pelo alvo em várias faixas do espectro eletromagnético (bandas)

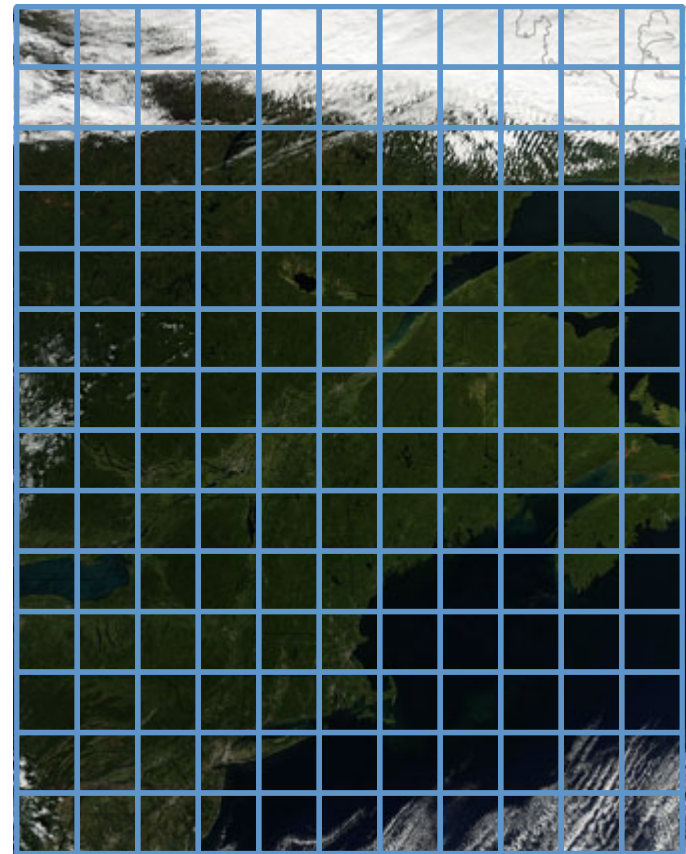


Cada alvo possui uma resposta espectral

O que é um Raster?

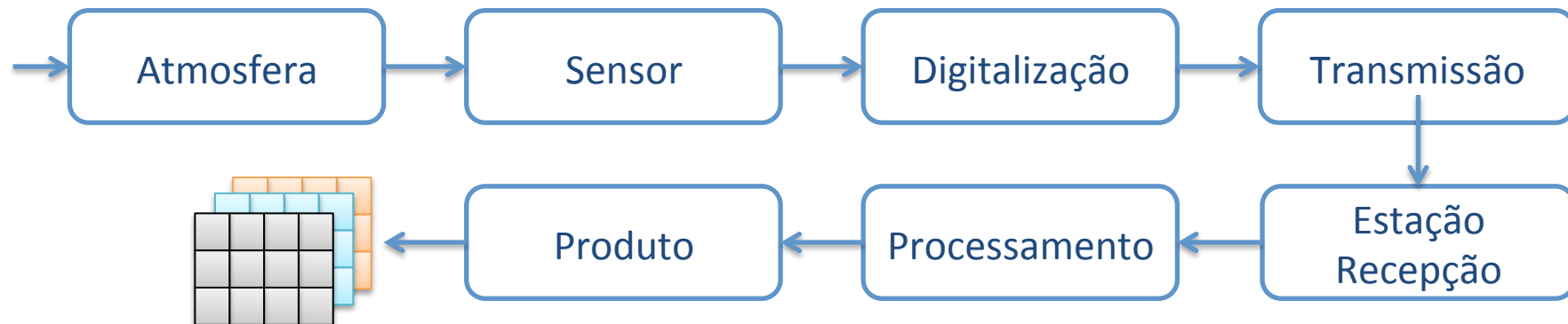
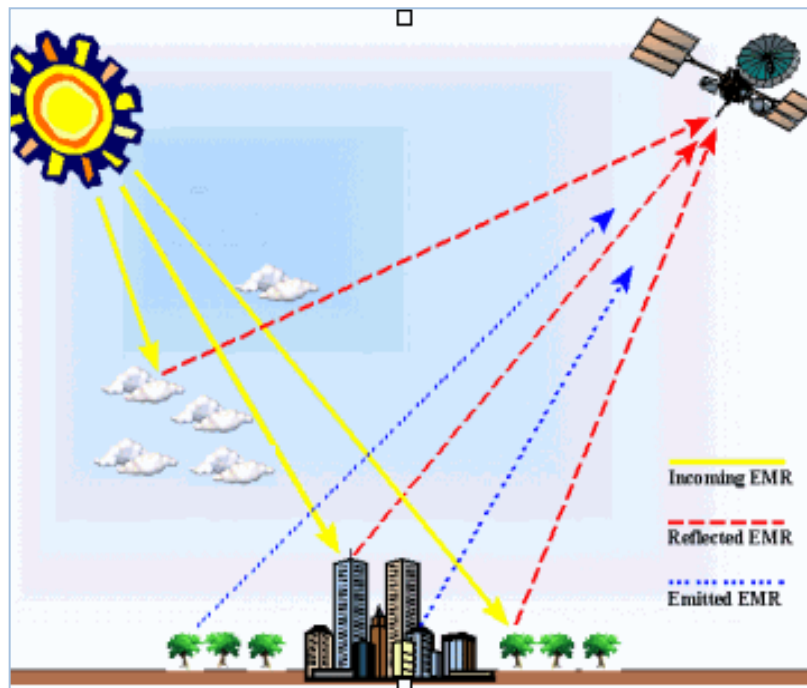
- Um raster ou dado matricial é uma matriz de células (ou pixels) associadas com localizações no espaço, formando uma grade regular.

September 26, 2013 - New England



Fonte: NASA, Imagem do dia: <http://modis.gsfc.nasa.gov/gallery/>

Coleta de Dados: Sensoriamento Remoto (Imagens Ópticas)



Fonte: Adaptado de Fonseca e Körting (2013)

Exemplo: Landsat (4-5) Thematic Mapper (TM)

Banda	Wavelength	Descrição	Características	Resolução
1	0.45-0.52	Visible (blue)	Penetração máxima na água Análise de vegetação x solo	30
2	0.52-0.60	Visible (green)	Vigor das plantas (pico de reflectância)	30
3	0.63-0.69	Visible (red)	Absorção de clorofila Discriminação de vegetação	30
4	0.76-0.90	Near infra-red	Biomassa de vegetação Identificação de culturas	30
5	1.55-1.75	Mid infra-red	Mapeamento vegetação Solo/vegetação	30
6	10.40-12.50	thermal	Temperatura; classificação de vegetação	120
7	2.08-2.35	Mid infra-red	Recursos minerais	30

scene size \approx 170 km north-south \times 183 km east-west (see [FAQ Landsat Missions](#))

Color composition => true colors (3, 2, 1) / highlight vegetation (5, 4, 3)

Fonte: Adaptado de Fonseca e Körting (2013)

Resolução Espacial



30 m



10 m

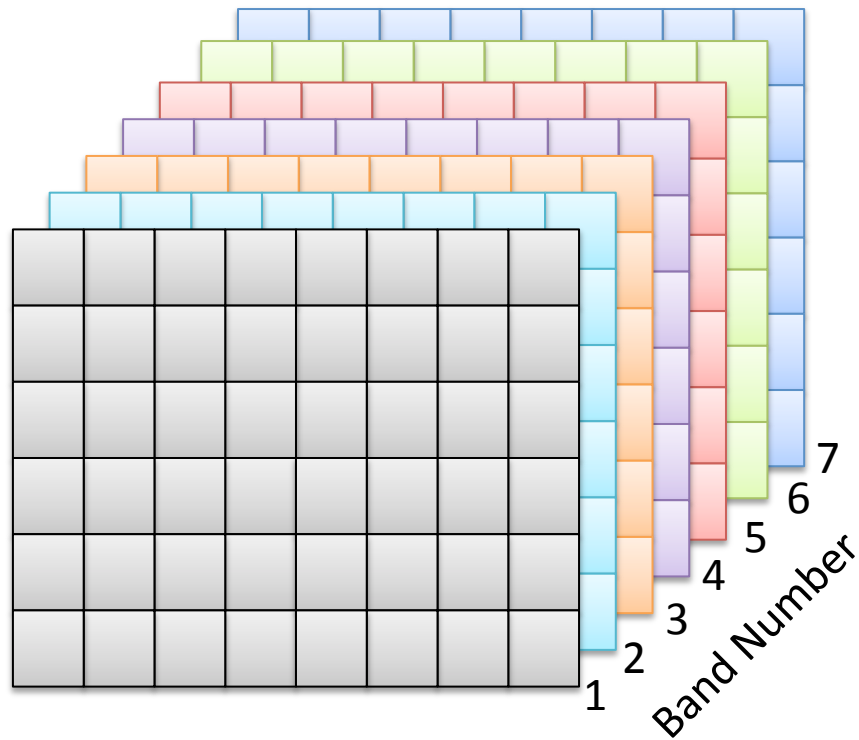


1 m

Satellite-Sensor	Spatial Resolution
SPOT-pan	10 m
SPOT-X	20 m
Landsat-5-TM	30 m
Landsat-MSS	80 m
AVHRR	1100 m

Fonte: Adaptado de Fonseca e Körting (2013)

Raster → Bandas

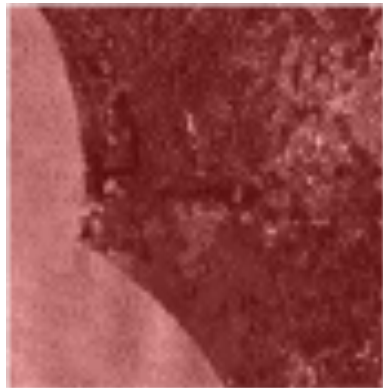


Considerações:

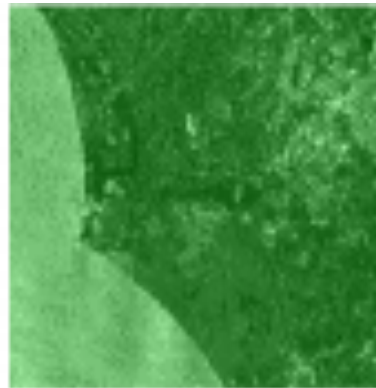
- Spectral Bands
- Scene size
- Spatial Resolution
- Temporal Resolution
- Quantized: 8 bits, 12 bits, 16 bits, ...
- Spatial Reference System

Raster → Bandas

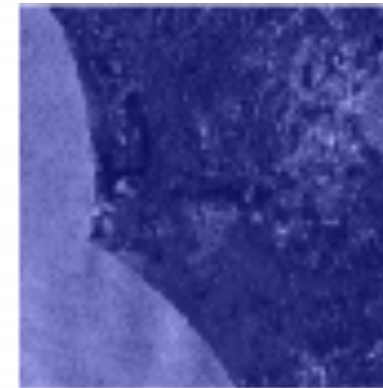
- Um arquivo raster por banda (ex: CBERS):
 - CBERS_2B_CCD1XS_20080223_148_120_L2_BAND2.tif (raster 1)
 - CBERS_2B_CCD1XS_20080223_148_120_L2_BAND3.tif (raster 0)
 - CBERS_2B_CCD1XS_20080223_148_120_L2_BAND4.tif (raster 2)



raster 0



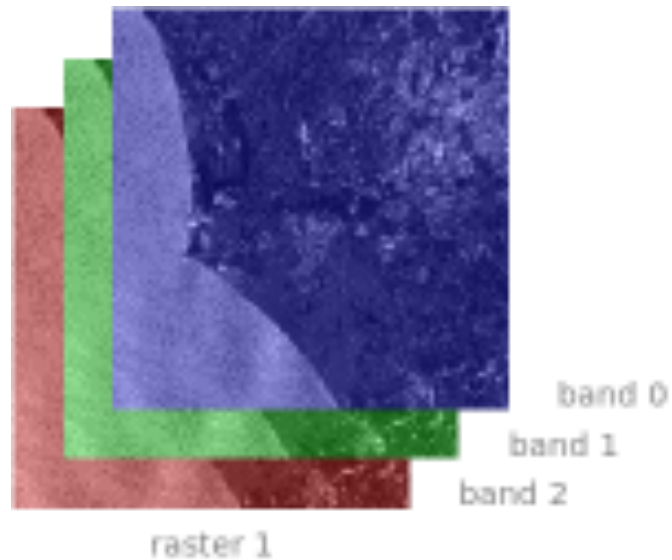
raster 1



raster 2

Raster → Bandas

- Um arquivo raster com múltiplas bandas (ex: World View)
 - IMG_BAND_2_3_4.tif



Raster → Bandas

- Cada banda pode estar associado a um tipo de dado.
- Metadados:
 - Descrição;
 - Autor;
 - Resolução;
 - Scale factor;
- Missing data.
- Block Width.
- Block Height.
- Pixel/band interleaved
- ...

*Discutir: imagens com palletete

Formatos de Dados Raster e APIs

- Hierarchical Data Format ([HDF](http://www.hdfgroup.org/)):
 - “HDF (Hierarchical Data Format) technologies are relevant when the data challenges being faced push the limits of what can be addressed by traditional database systems, XML documents, or in-house data formats...” see: http://www.hdfgroup.org/why_hdf
 - Target: **array-oriented scientific data**
 - machine-independent data format, self-describing, compression, chunking
- Network Common Data Form ([NetCDF](http://www.cgd.cornell.edu/netcdf/)):
 - Target: **array-oriented scientific data**
 - self-describing, machine-independent data format
- GeoTIFF:
 - TIFF, BigTIFF and GeoTIFF
 - TIFF tags with metadata
 - Tiled, Pixel/Band Interleaved, Compression

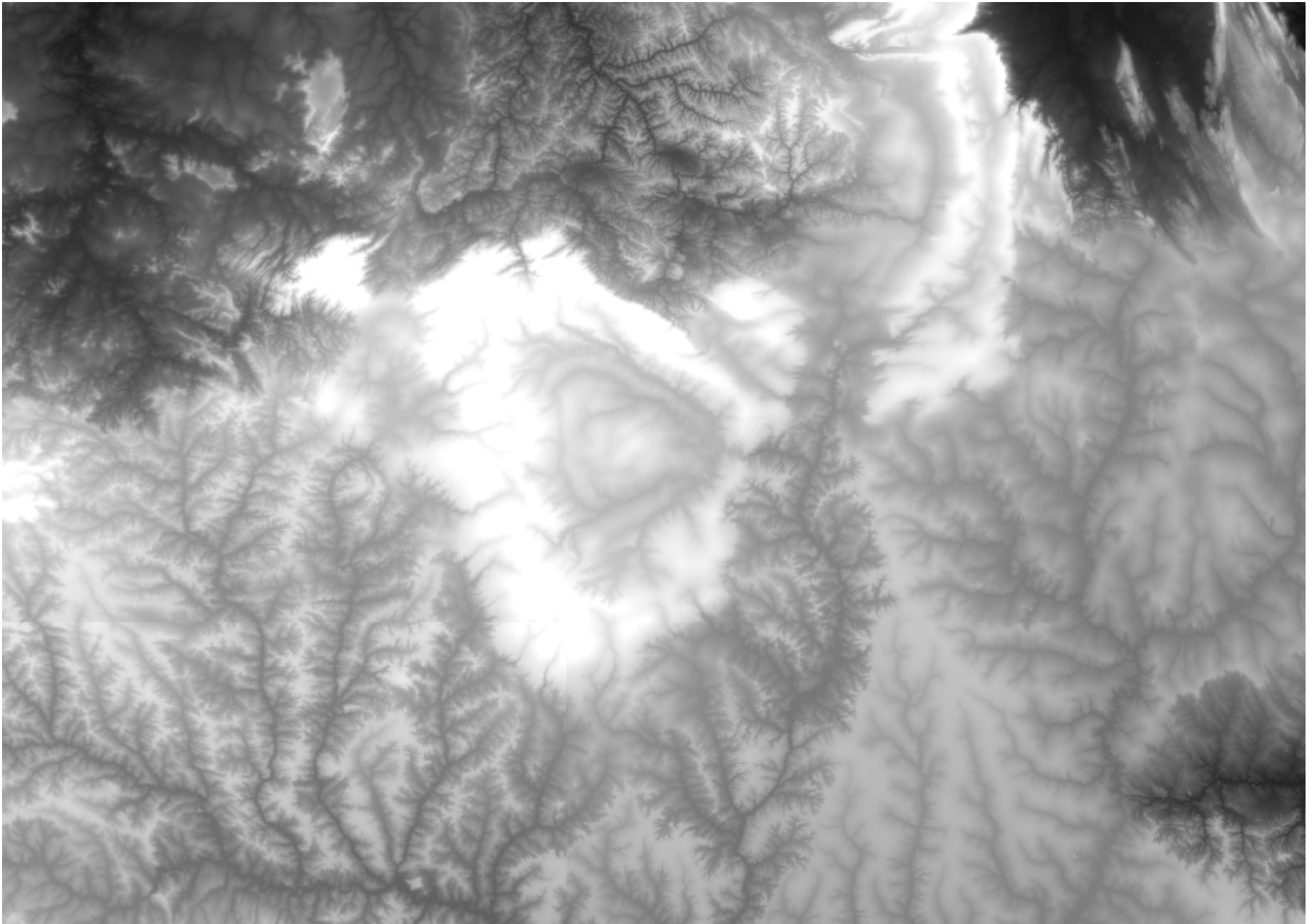
See http://www.gdal.org/formats_list.html for a long list of raster data formats

Hands-on

Visualização de imagens no QGIS



Imagem Colorida – Cena CBERS-2B



Mosaico SRTM – 4 Cartas 1:250.000



Composição Colorida – Cena LandSat L71221071_07120060531
True Color - Bandas: 3, 2, 1



Composição Colorida – Cena Landsat L71221071_07120060531
Realce da Vegetação - Bandas: 5, 4, 3

Dica: usar `gdalinfo` para obter informações de uma imagem



\$ gdalinfo bahia.tif

Estudo de Caso: dados do sensor MODIS

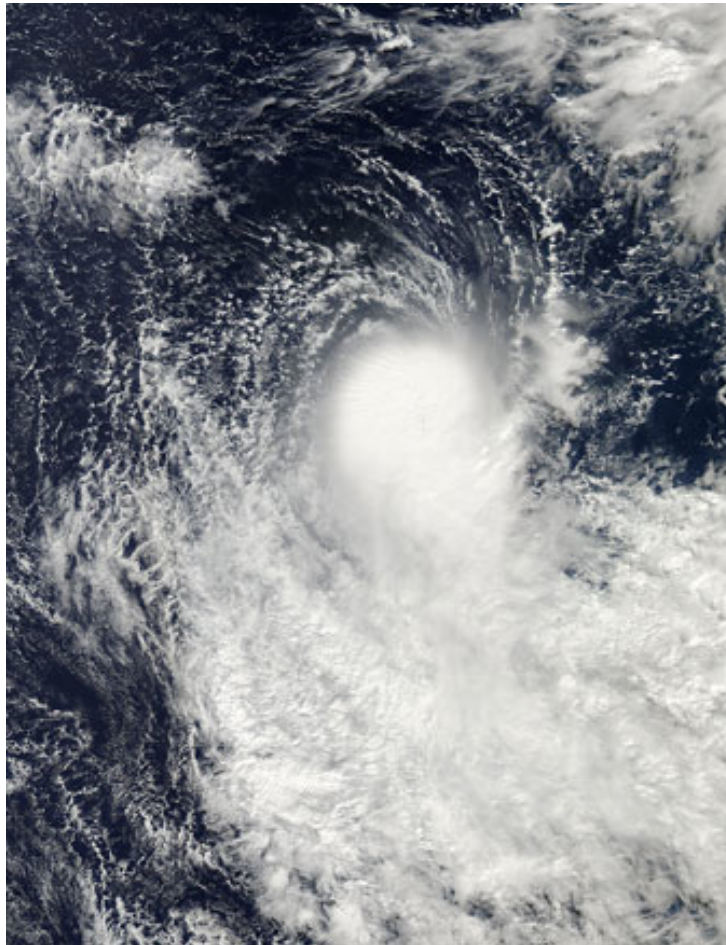
O que é o MODIS?

Tiles + Datasets + Multi-Spectra Bands = size

Frequência Dados?

Big data e MODIS (Volume + Velocidade)?

Sensor MODIS: Moderate Resolution Imaging Spectroradiometer



Tropical Cyclone Jack (24S) in the South Indian Ocean

Satellite: Aqua
Date Acquired: 4/21/2014
Resolutions: 500m
250m
Bands Used: 1,4,3
Credit: Jeff Schmaltz (MODIS Land Rapid Response Team, NASA GSFC)



Terra (EOS AM-1)

Source: [Wikipedia](http://en.wikipedia.org/wiki/Terra_(satellite)) (April, 2014)



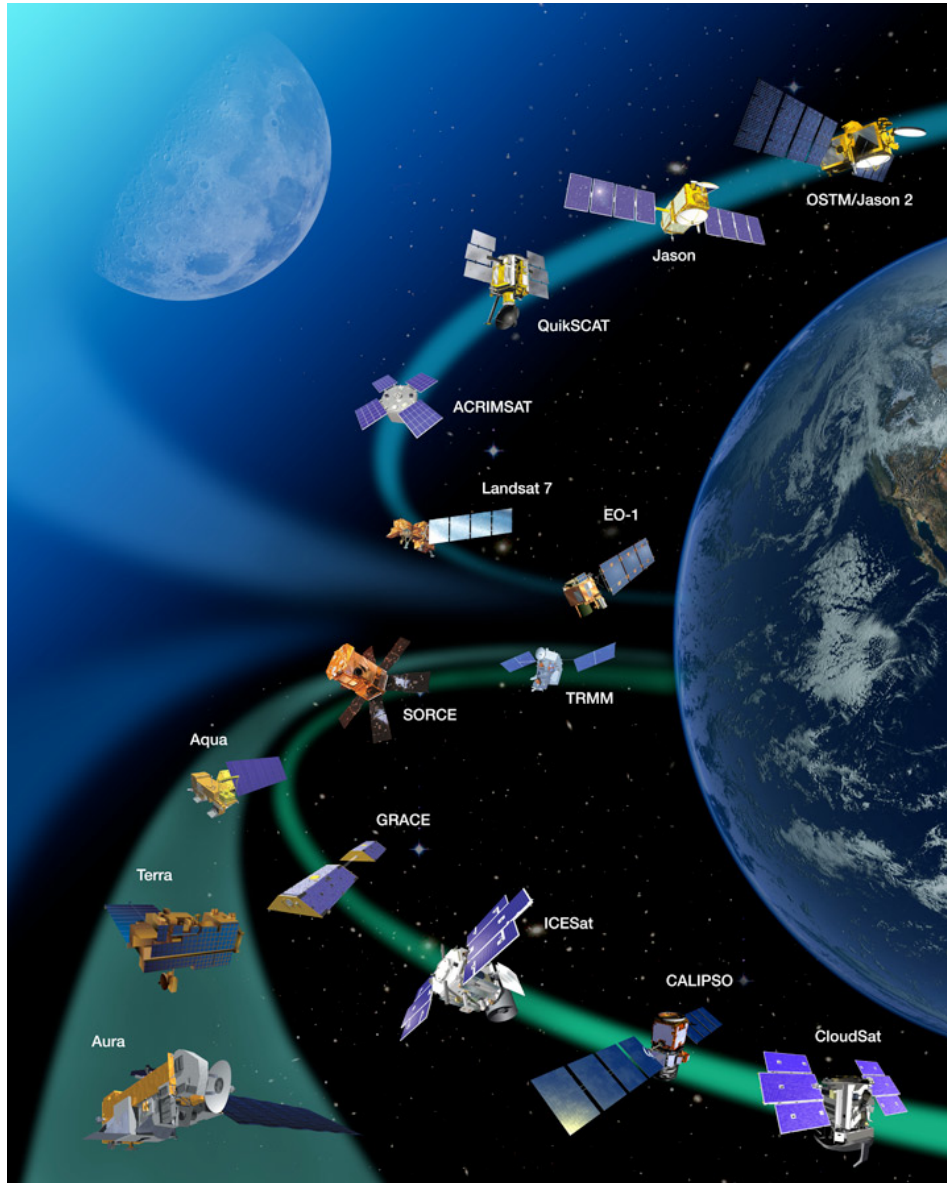
Aqua (EOS PM-1)

Source: [Wikipedia](http://en.wikipedia.org/wiki/Aqua_(satellite)) (April, 2014)

Source: <http://modis.gsfc.nasa.gov> (April, 2014)

NASA's EOS (Earth Observing System)

Fonte: [NASA](#) (April, 2014)



Atmosphere



Land



Ocean

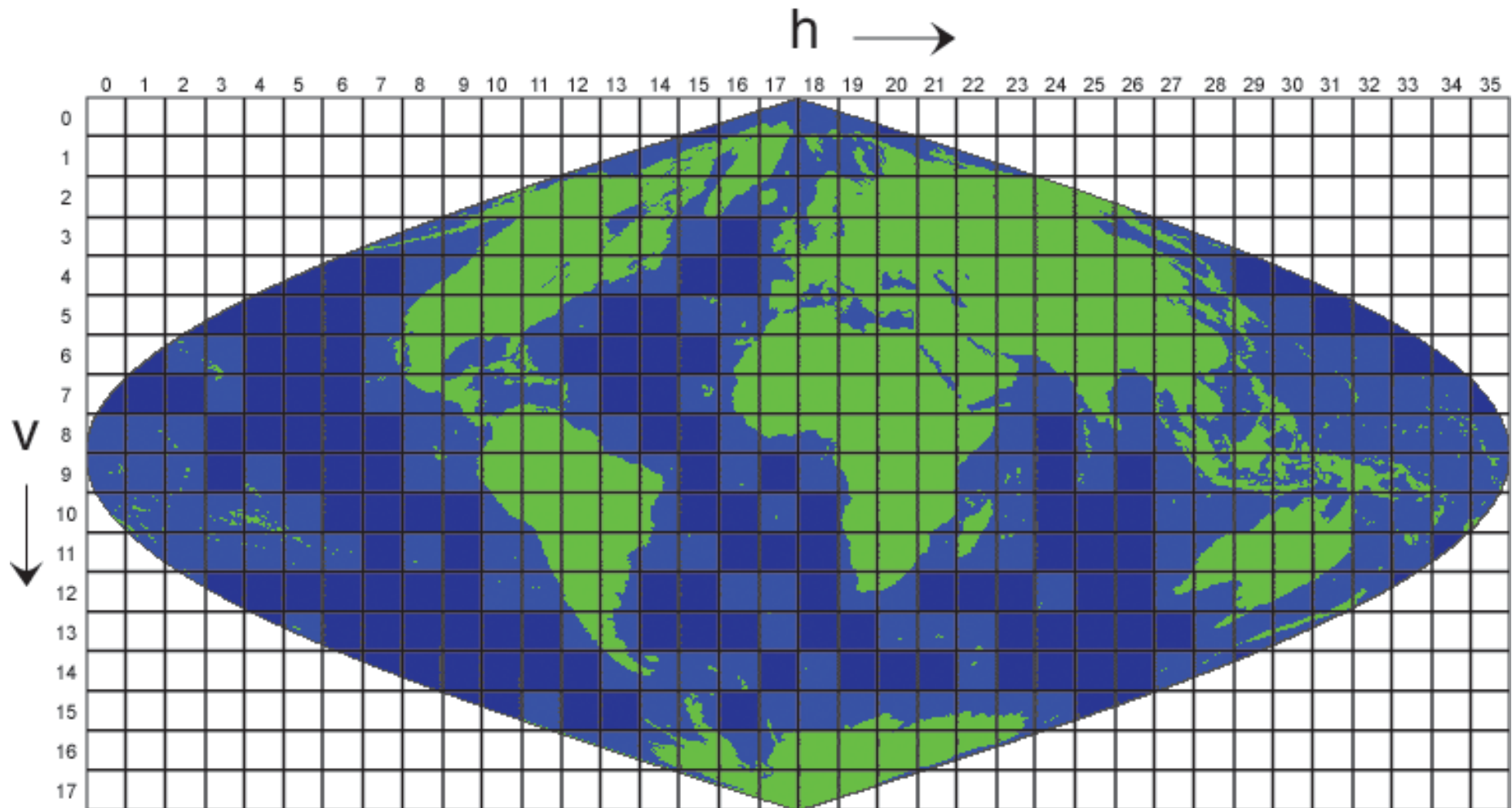


Calibration

Vamos conferir como é organizado um
produto MODIS:

MOD13Q1 (Terra)

MODIS: Grade Sinusoidal



Fonte: http://nsidc.org/data/modis/data_summaries/landgrid.html

Arquivos Originais → HDF

- Cada arquivo contém 12 *subdatasets* (ou bandas):
 - 11 bandas de 16-bits;
 - 1 banda de 8-bits (*pixel reliability*).
- Arquivos de tamanho variável: ~5 MiB → ~230 MiB
- Cada tile (cena) possui 4.800 x 4.800 pixels.
- Internamente os HDFs são organizados em blocos de 4.800 x 208 pixels.
- CRS:
`+proj=sinu +lon_0=0 +x_0=0 +y_0=0 +a=6371007.181 +b=6371007.181
+units=m +no_defs`
- Resolução Temporal: 16-dias

Brasil → 30 Tiles

- Considerando 15 anos de observação.
- Para cada tile: $23 * 15 \rightarrow 345$ imagens.
- 30 tiles: $30 * 345 \rightarrow 10.350$ imagens.
- 1 imagem $\rightarrow 4800 \times 4800 \rightarrow 23.040.000$ pixels.
- $T_0 \rightarrow 30$ -tiles $\rightarrow 691.200.000$ pixels.
- 10.350 imagens: 238.464.000.000 pixels.
- Considerando apenas o atributo (banda) red-reflectance:
 $238.464.000.000 \times 2 \text{ bytes} = 444.17 \text{ GiB}$

Exemplo: Mapa de Vegetação Global de Fevereiro de 2000 a Fevereiro de 2014



Fonte: [NASA Earth Observatory](#) (23 de Abril , 2014)

Exemplo: Time Series in Land Use and Land Cover Change (LUCC) Studies

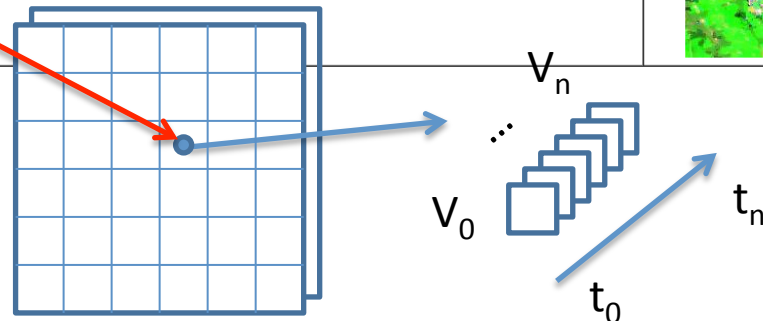
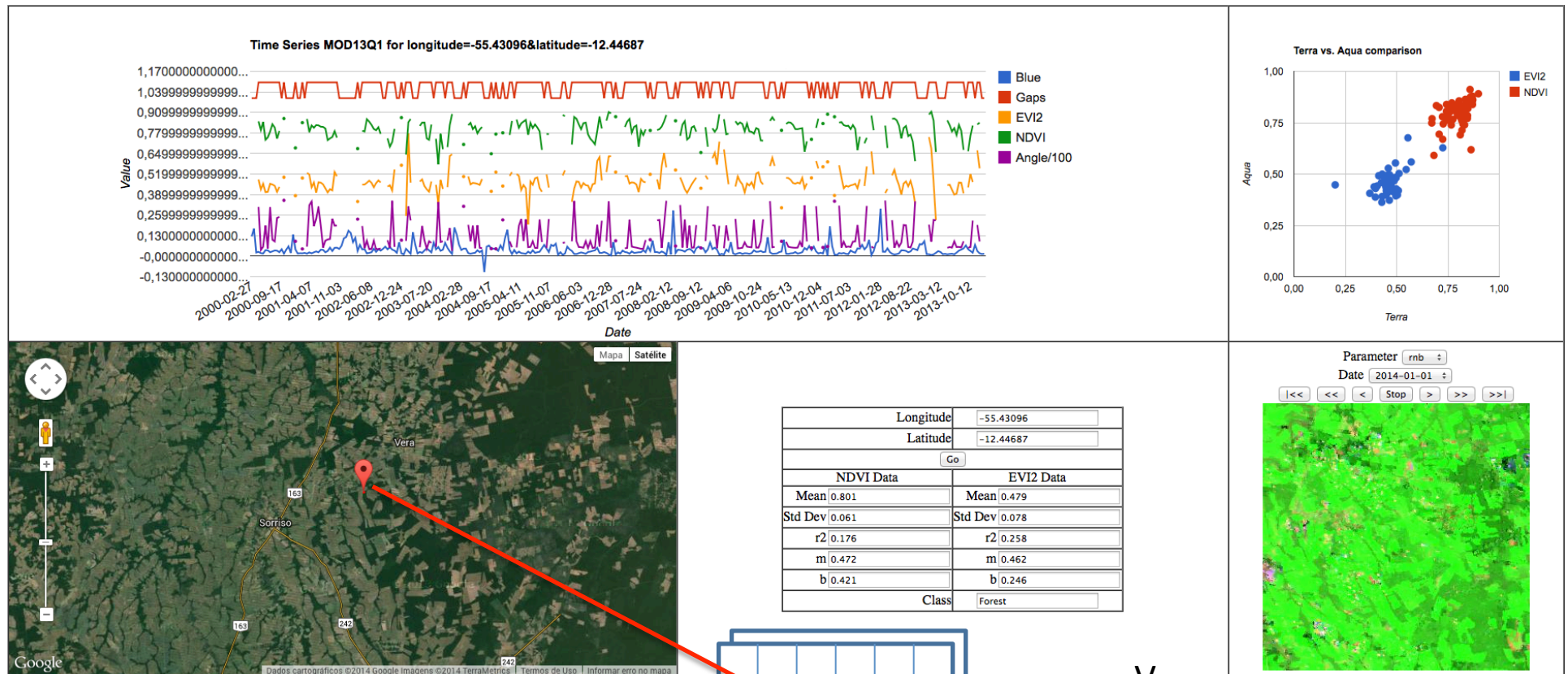


Tabela de Produtos MODIS

https://lpdaac.usgs.gov/products/modis_products_table

<http://modis.gsfc.nasa.gov/data/dataproduct>

Observação: Neste curso não iremos abordar imagens de radar.

Discussão: Como Armazenar e Gerenciar
Dados Matriciais?

Solução 1: Usando um SGBD-R com
suporte matricial

PostGIS Raster

Oracle GeoRaster

Solução 1: Usando um SGBD-R com suporte matricial

PostGIS Raster



Oracle GeoRaster

PostGIS Raster

Raster (Matricial - Células Multibandas)

Visão do espaço na forma de uma grade retangular,
com células contendo uma ou mais valores numéricos

PostGIS Raster

- O projeto do PostGIS Raster possibilita trabalhar com vários casos de uso com imagens:
 - Armazém de imagens (possivelmente não relacionados)
 - Tiles:
 - Regulares ou irregulares (pode ter missing tiles)
 - *in db x out db storage*
- Novas visões com metadados das imagens:
 - raster_columns
 - raster_overviews

PostGIS Raster → SQL

- Table creation:

```
CREATE TABLE raster_table  
( rid SERIAL PRIMARY KEY,  
  rast RASTER  
);
```

- Spatial index creation:

```
CREATE INDEX spidx_table_col ON raster_table  
  USING gist(ST_ConvexHull(raster-col));
```

Organizando uma imagem em blocos (Tiles)

- Uma estratégia muito comum dos SIG é particionar uma imagem em blocos durante o armazenamento.
- O PostGIS raster suporta esta estratégia de armazenamento.

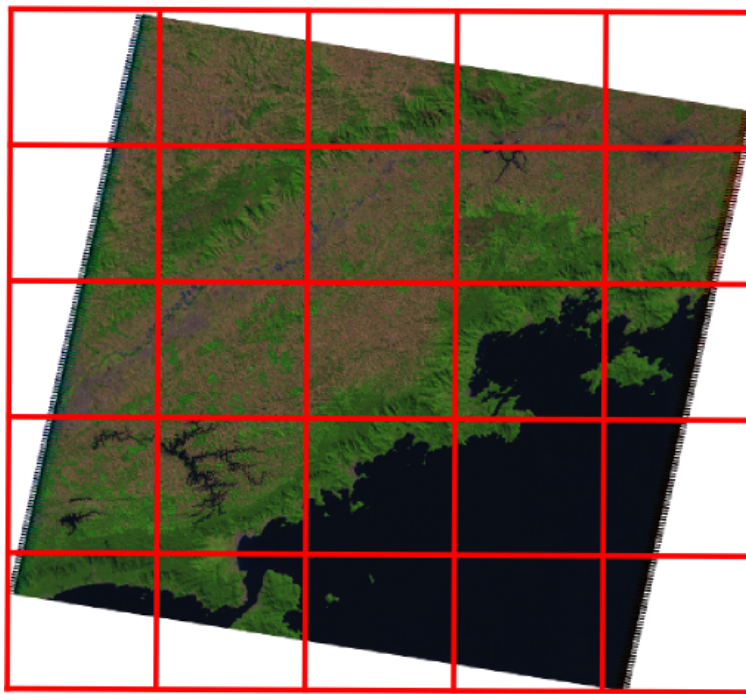




Tabela: imagem_sp	
gid	raster_tile
1	
2	
...	

Vamos carregar um raster para o banco

raster2pgsql

raster2pgsql

- Ao executar o comando:
`$ raster2pgsql -?`
- recebemos uma mensagem informando a versão da GDAL:
`RELEASE: 2.1.6 GDAL_VERSION=111 (r13384)`
- A sintaxe básica deste comando é:
`raster2pgsql [<options>] <raster>[<raster>[...]]
[[<schema>.]<table>]`
- Múltiplos rasters podem ser especificados usando: *, ?.

raster2pgsql: Opções

Parâmetro	Descrição	Opicional
-s <srid>	Define o SRID do raster (default: 0).	✓
-b <band>	Lista separada por vírgula das bandas a serem extraídas de um raster. A primeira banda é numerada a partir de 1. Pode-se usar hífen (-) para definir faixas de bandas. Caso este parâmetro não seja informado, todas as bandas do raster serão extraídas.	✓
-t <tile size>	Define o tamanho de <i>tile</i> ("WIDTHxHEIGHT"). Pode-se utilizar o valor "auto" para se obter um tamanho automático.	X
-P	Faz o pad dos tiles right-most e bottom-most para garantir que todos os tiles tenham a mesma largura e altura.	✓
-R	Registra o raster como sendo armazenado fora do banco.	✓
-f <column>	Nome da coluna raster.	✓

raster2pgsql: Opções

Parâmetro	Descrição	Opicional
-d	Remove a tabela, cria e faz a carga com os dados do raster.	✓
-a	"Apenda" o raster em uma tabela existente que deve ter um esquema compatível para esta operação (mesmo esquema).	✓
-c	Cria uma nova tabela e faz a carga (opção padrão).	✓
-p	Prepare mode (apenas cria a tabela).	✓
-l <overview factor>	Lista separada por vírgula dos fatores de criação de overviews do raster.	✓
-I	Cria um índice espacial sobre a coluna raster.	✓
-N <nodata>	Valor de NODATA para as bandas que não possuem informações de NODATA.	✓
-G	Lista os formatos de imagens suportados pela GDAL.	✓

raster2pgsql: Opções

Parâmetro	Descrição	Opicional
-C	Habilita a criação de um conjunto de restrições padrões sobre a coluna raster após a carga dos dados.	✓
-x	Evita que a restrição de "max extent" seja incluída.	✓
-r	Habilita restrições para blocagem regular.	✓
-M	Roda VACUUM ANALYZE para a tabela com a coluna raster.	✓
-F	Adiciona uma coluna com o nome do arquivo raster.	✓
-n <column>	Especifica o nome da coluna que conterà o nome do arquivo raster.	✓
-T <tablespace>	Especifica a tablespace da nova tabela.	✓
-X <tablespace>	Especifica a tablespace do novo índice criado.	✓
-k	Não faz o check de valores NODATA para as bandas.	✓

```
$ raster2pgsql -G
```

Na minha instalação os seguintes formatos de imagem são suportados para leitura pela ferramenta raster2pgsql

Formatos Raster Suportados pela GDAL

ACE2	Convair PolGASP	Envisat Image Format
ARC Digitized Raster Graphics	DIPEX	Erdas .LAN/.GIS
ASCII Gridded XYZ	DRDC COASP SAR Processor Raster	Erdas Imagine Images (.img)
AirSAR Polarimetric Image	DTED Elevation Raster	Erdas Imagine Raw
Arc/Info ASCII Grid	ECRG TOC format	FARSITE v.4 Landscape File (.lcp)
Arc/Info Binary Grid	ELAS	FIT Image
Arc/Info Export E00 GRID	ENVI .hdr Labelled	Fuji BAS Scanner Image
Azavea Raster Grid format	EOSAT FAST Format	GRASS ASCII Grid
CEOS Image	ERMMapper .ers Labelled	GRidded Binary (.grb)
CEOS SAR Image	ESRI .hdr Labelled	GSC Geogrid
COSAR Annotated Binary Matrix (TerraSAR-X)	EUMETSAT Archive native (.nat)	Generic Binary (.hdr Labelled)
CTable2 Datum Grid Shift	EarthWatch .TIL	...

Importando a imagem bahia.tif

```
$ raster2pgsql -c -C -P -r -s 31984 -I -M  
-t 128x128 bahia.tif  
public.img_bahia > img_bahia.sql
```

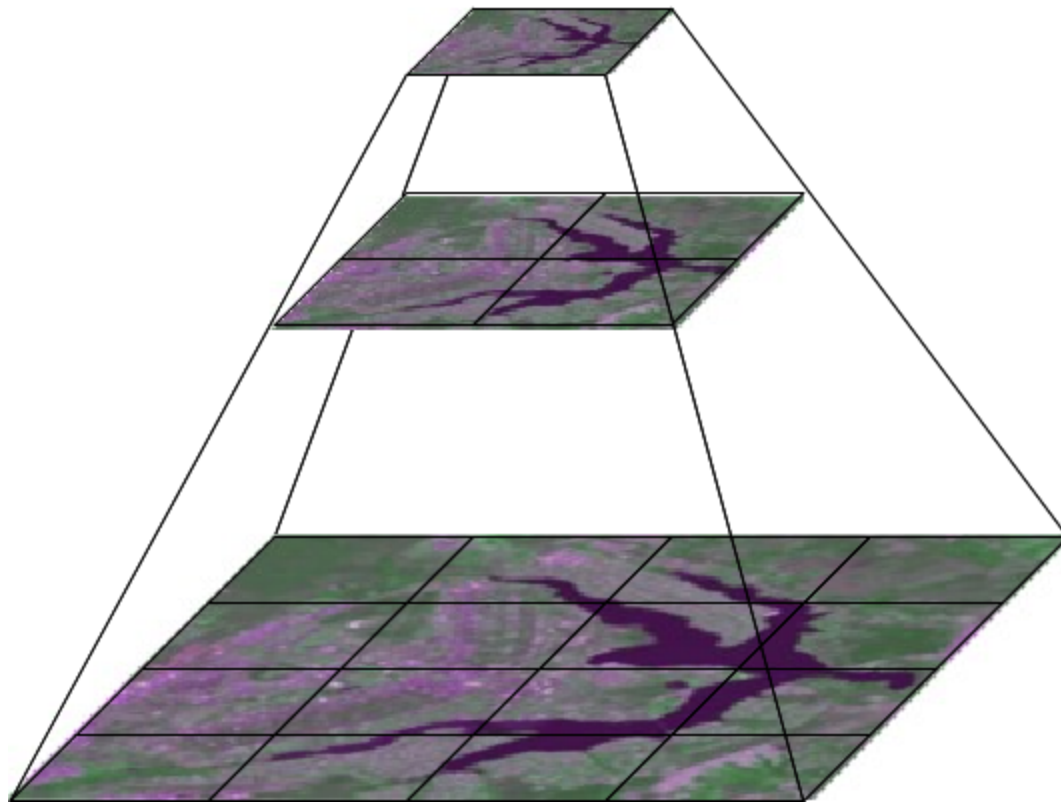
- Onde:
 - s 31984
 - c: criar uma nova tabela e populá-la;
 - C: aplicar algumas constraints;
 - r: aplicar constraints de blocagem regular;
 - t: tile size;
 - I: criar índice espacial sobre a coluna raster;
 - M: executar vacuum analyze na tabela raster.
- Importar para o banco:

```
$ psql -U postgres -d bdgcurso -f img_bahia.sql
```

Visualizando tabelas raster com o Quantum GIS

Overviews (Pirâmides/Multi-resolução)

- Parâmetro raster2pgsql: -1 2,4



$$R_2 = R \times 2^2$$

$$R_1 = R \times 2^1$$

$$R_0 = R \times 2^0$$

Importando imagens com multiresolução

```
$ raster2pgsql -c -C -P -r -s 31984 -I -M  
-t 128x128 -l 2,4,8,16 bahia.tif  
public.img_bahia_mr > img_bahia_mr.sql
```

- Importar para o banco:

```
$ psql -U postgres -d bdgcurso -f img_bahia_mr.sql
```

Exercícios - Raster

1. Importar as imagens abaixo para o PostgreSQL e visualizá-las no QGIS:
 - CBERS-2B (SRID: 29183):
 - Fusao_CBERS.tif
 - LandSat (SRID: 32623):
 - L71221071_07120060531_B10.TIF
 - L72221071_07120060531_B20.TIF
 - L71221071_07120060531_B30.TIF
 - SRTM (SRID: 4326):
 - 15_48_ZN.tif
 - 15_495ZN.tif
 - 16_48_ZN.tif
 - 16_495ZN.tif

Processamento de Imagens no SGBD

Discussão: Atualmente, não há um padrão bem definido para as extensões matriciais.

Discussão: Existe algum problema com o projeto do suporte Raster dos atuais SGBD-R?

Como lidar com os requisitos de aplicações de EO que podem necessitar como entrada dados massivos?


Solução 2: Usando um Array Database

rasdaman

SciDB

Solução 2: Usando um Array Database

rasdaman

SciDB 

Novas Tecnologias de Bancos de Dados para Dados Matriciais

Array Databases

“Arrays as first class citizens”

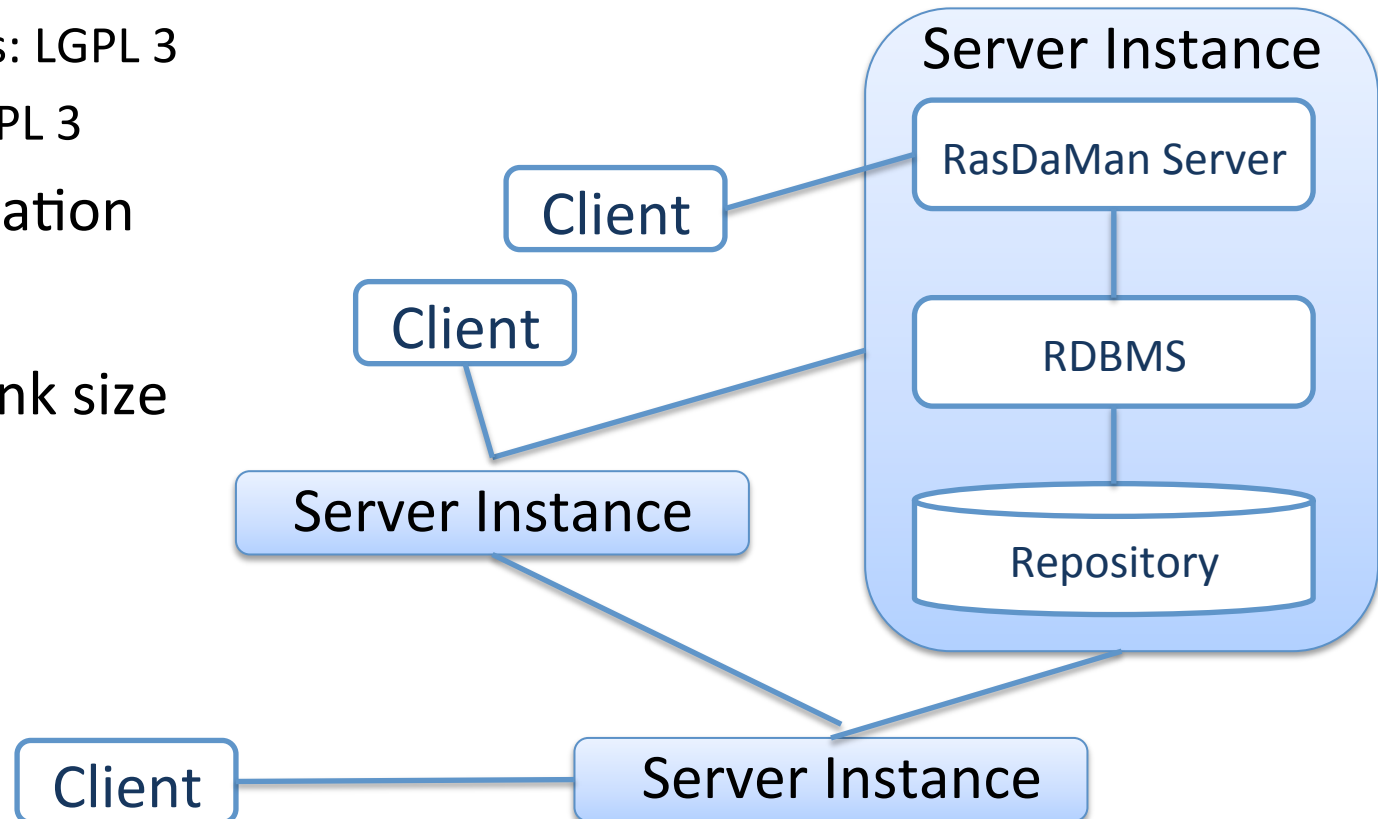
Raster Data Manager (rasdaman)

Prof. Peter Baumann

Jacobs University, Computer Science, Bremen,
Germany

rasdaman

- Domain-neutral Array Database System.
- Site: <http://www.rasdaman.org>
- License:
 - Client libraries: LGPL 3
 - Server side: GPL 3
- Server Federation
- OGC WCPS
- Variable chunk size



rasdaman → rasql

- Array Algebra → Embedded Query Language (SQL) → rasql
- Example:
 - * “subtracting each MDD of collection mr2 from each MDD of collection mr where at least one difference pixel value is greater than 50”:

```
select mr - mr2
  from mr, mr2
 where some_cells( mr - mr2 > 50 )
```

SciDB

“SciDB is an open-source analytical database oriented toward the data management needs of scientists.”

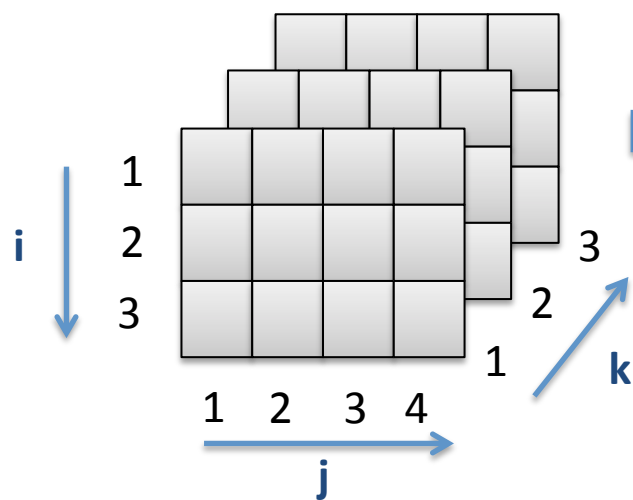
(Stonebraker et al., 2013)

Arrays, Dimensões, Atributos e Particionamento

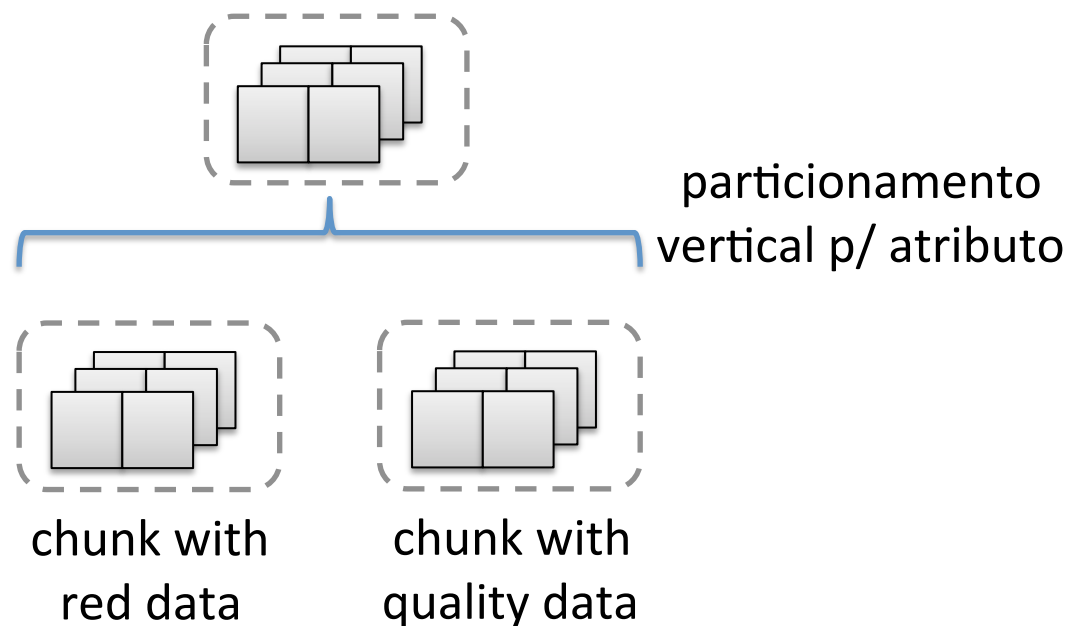
mod09q1

<attributes>

[j=1:4, 2, 1, i=1:3, 1, 1, t=1:*, 3, 0]



Array é dividido em chunks
chunk-size: 2 x 1 x 3



- ✓ Replicação
- ✓ Compressão
- ✓ Versionamento

Linguagens de Consulta: AQL e AFL

```
AQL% SELECT sqrt(v1) INTO a3  
      FROM a1  
      WHERE j = 1;
```

```
AQL% SELECT sum(v1) FROM a5 WHERE j = 1;
```

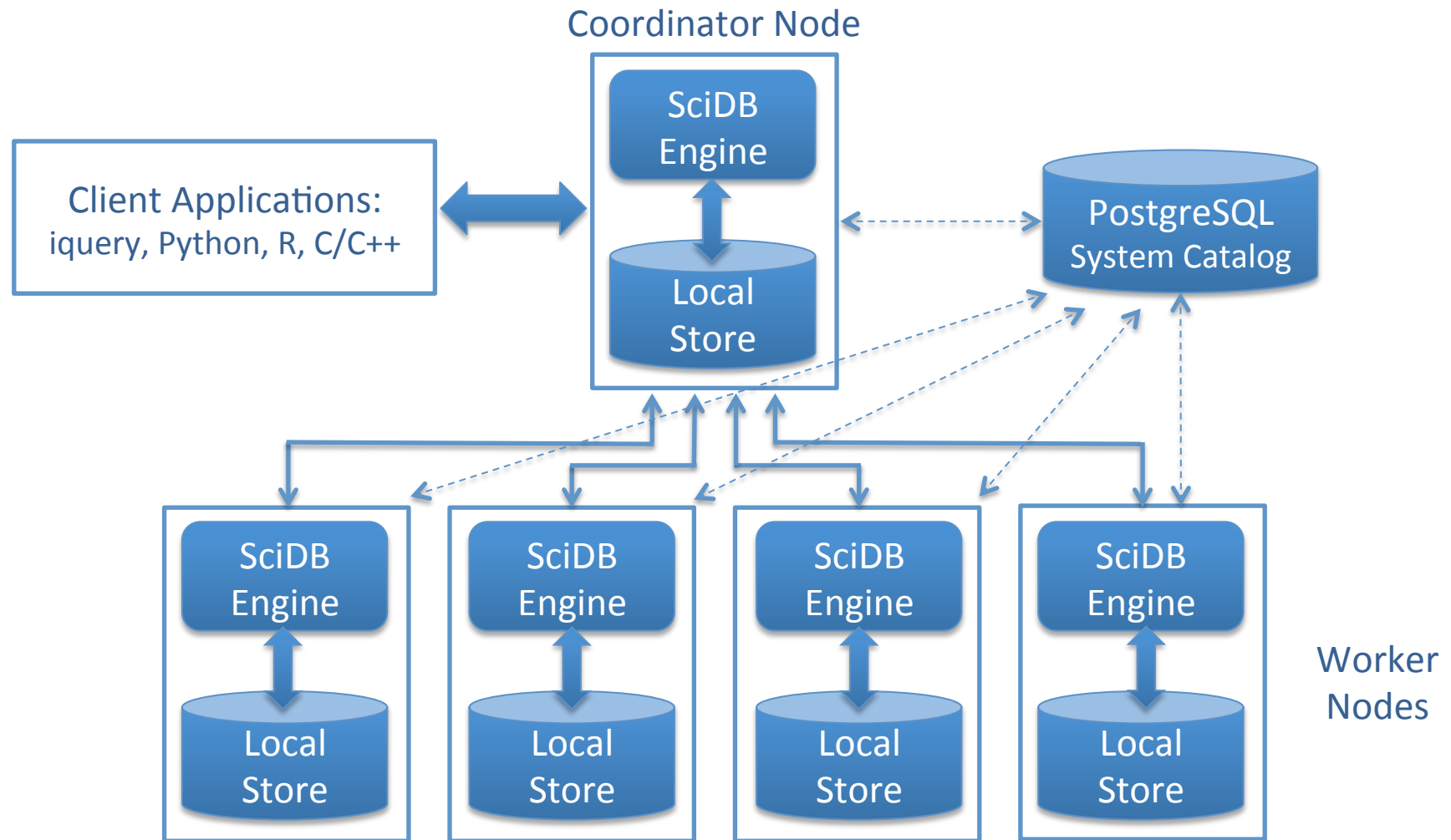
```
AFL% build(<v2:double>[j=1:4,2,1, y=1:4,2,1],  
          j - i);
```

1	0	1	2	3
2	-1	0	1	2
3	-2	-1	0	1
4	-3	-2	-1	0
	1	2	3	4

Integração com ambientes estatísticos e de álgebra linear

- R
- [ScaLAPACK](#) — Scalable Linear Algebra PACKage

Arquitetura



Source: Adapted from PARADIGM4

Behind SciDB

- Founders: Mike Stonebraker, Andy Palmer, David DeWitt, Kian-Tat Lim, Jacek Becla
Science Advisory Board:
 - “... we have spent a large amount of time talking to scientists about their requirements from a data management system (see the “Use Cases” section of the scidb.org website) ...” Seering et al. (2012).
- Large Synoptic Survey Telescope (LSST) project:
 - [Extreme Large Databases](#) (XLDB)
- Intel Science and Technology Center for Big Data:
<http://istc-bigdata.org>
- The develop team:
Company: Paradigm4

O que não se tem no SciDB?

- Loader de matrizes multidimensionais.
- As matrizes não tem um sistema de referencia espacial associado.
- As células não possuem informações de resolução.
- Não existem operadores espaciais como clipping c/ uma máscara vetorial.
- Não se tem um suporte a definição de metadados dos arrays. (Embora o catálogo esteja no PostgreSQL!)

Referências

Livros

- John R. Jensen. *Remote Sensing of the Environment*.

Artigos

- STONEBRAKER, M.; BROWN, P.; POLIAKOV, A.; RAMAN, S. ***The architecture of SciDB***. In Proceedings of the 23rd international conference on Scientific and statistical database management (SSDBM'11), Judith Bayard Cushing, James French, and Shawn Bowers (Eds.). Springer-Verlag, Berlin, Heidelberg, 2011, 1-16.
- TAFT, R.; VARTAK, M.; SATISH, N. R.; SUNDARAM, N.; MADDEN, S.; STONEBRAKER, M. ***Genbase: a complex analytics genomics benchmark***. Computer Science and Artificial Intelligence Laboratory Technical Report, MIT-CSAIL-TR-2013-028, November 19, 2013.

Cursos

- FONSECA, L. M. G.; KÖRTING, T. S. **PDI: Fundamentos**. Notas de Aula SER 437 - Processamento Digital de Imagens de Sensores Remotos. Disponível em: <http://www.dpi.inpe.br/cursos/ser437>. Acesso: 02 de Maio de 2014.

Videos

- [THAKAR, A](#) (Johns Hopkins University). **Billions of Stars on Off-the-shelf Software / The SDSS SkyServer and Beyond: Why "The Not-So-Little-Engine that Could" is Still Chugging Along.** 7th Extremely Large Databases Conference, September 9-12, 2013, Stanford University, California, USA. [Video](#), [Slides](#), Access: 04th April, 2014.