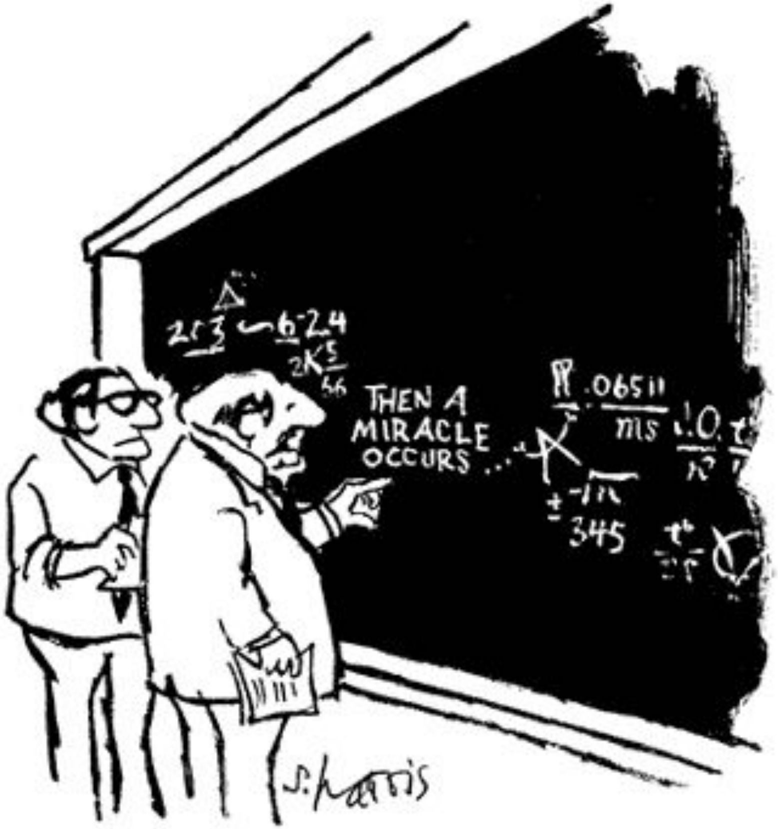
An exercise on reproducible science using array databases (SciDB)

alber sánchez

alber.ipia@inpe.br

Referata GeoInformática - 2015.05.19

Reproducible research

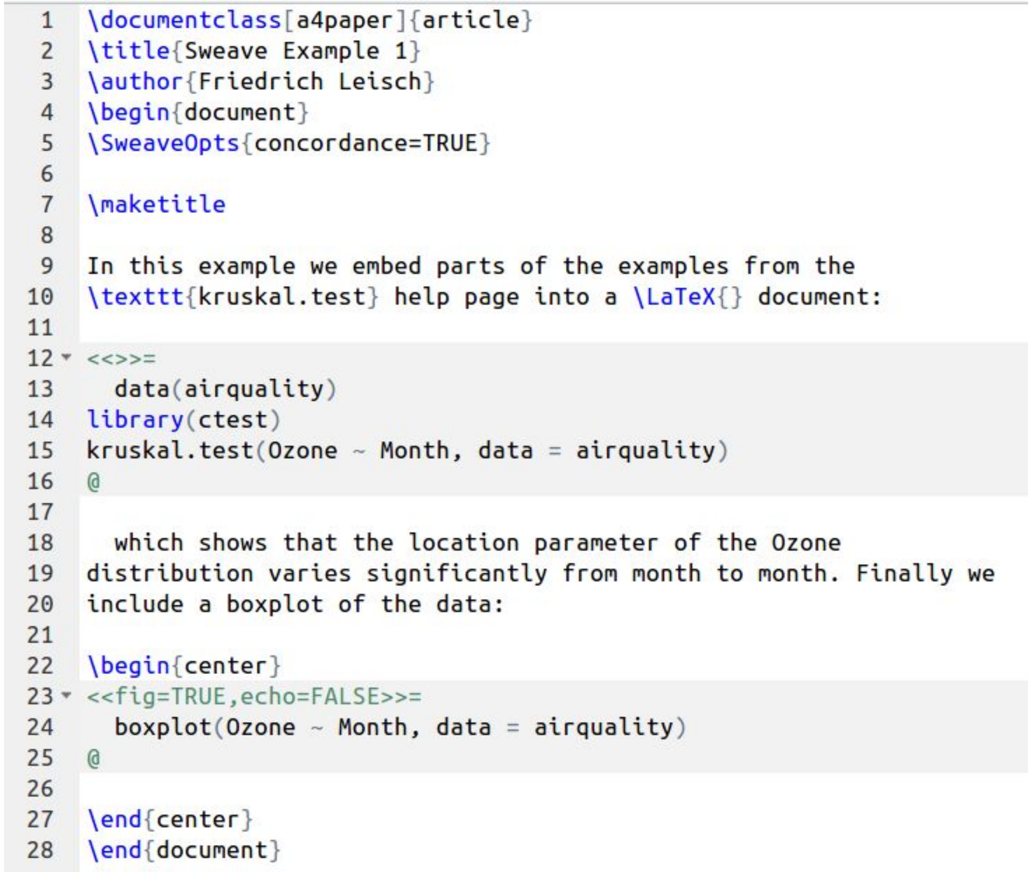


"I think you should be more explicit here in step two."

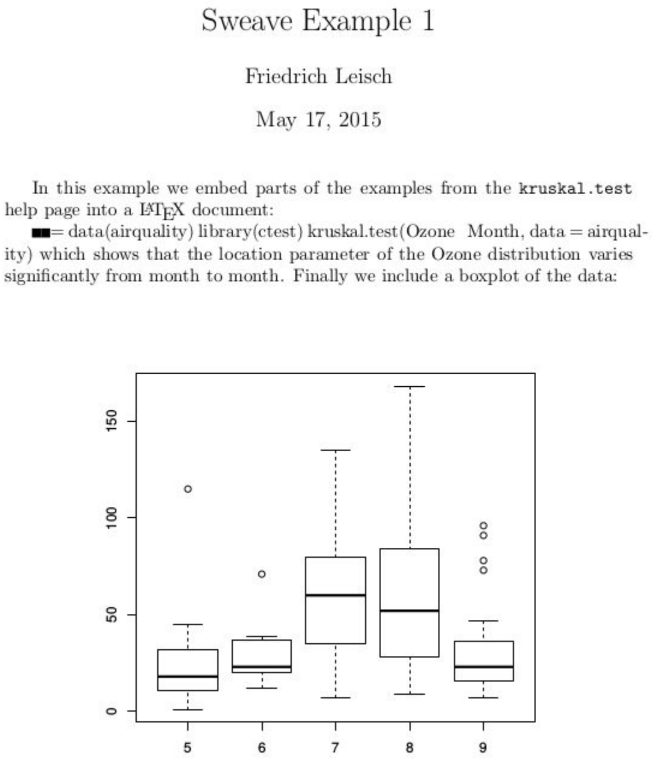
Source: http://www.sciencecartoonsplus.com/

Anything in a scientific paper should be reproducible by the reader.

Literate programming



```
\documentclass[a4paper]{article}
\title{Sweave Example 1}
\author{Friedrich Leisch}
\begin{document}
\SweaveOpts{concordance=TRUE}

\maketitle

In this example we embed parts of the examples from the
\texttt{kruskal.test} help page into a \LaTeX{} document:

<<>>=
  data(airquality)
library(ctest)
kruskal.test(Ozone ~ Month, data = airquality)
@

  which shows that the location parameter of the Ozone
distribution varies significantly from month to month. Finally we
include a boxplot of the data:

\begin{center}
<<fig=TRUE,echo=FALSE>>=
  boxplot(Ozone ~ Month, data = airquality)
@

\end{center}
\end{document}
```



Sweave Example 1

Friedrich Leisch

May 17, 2015

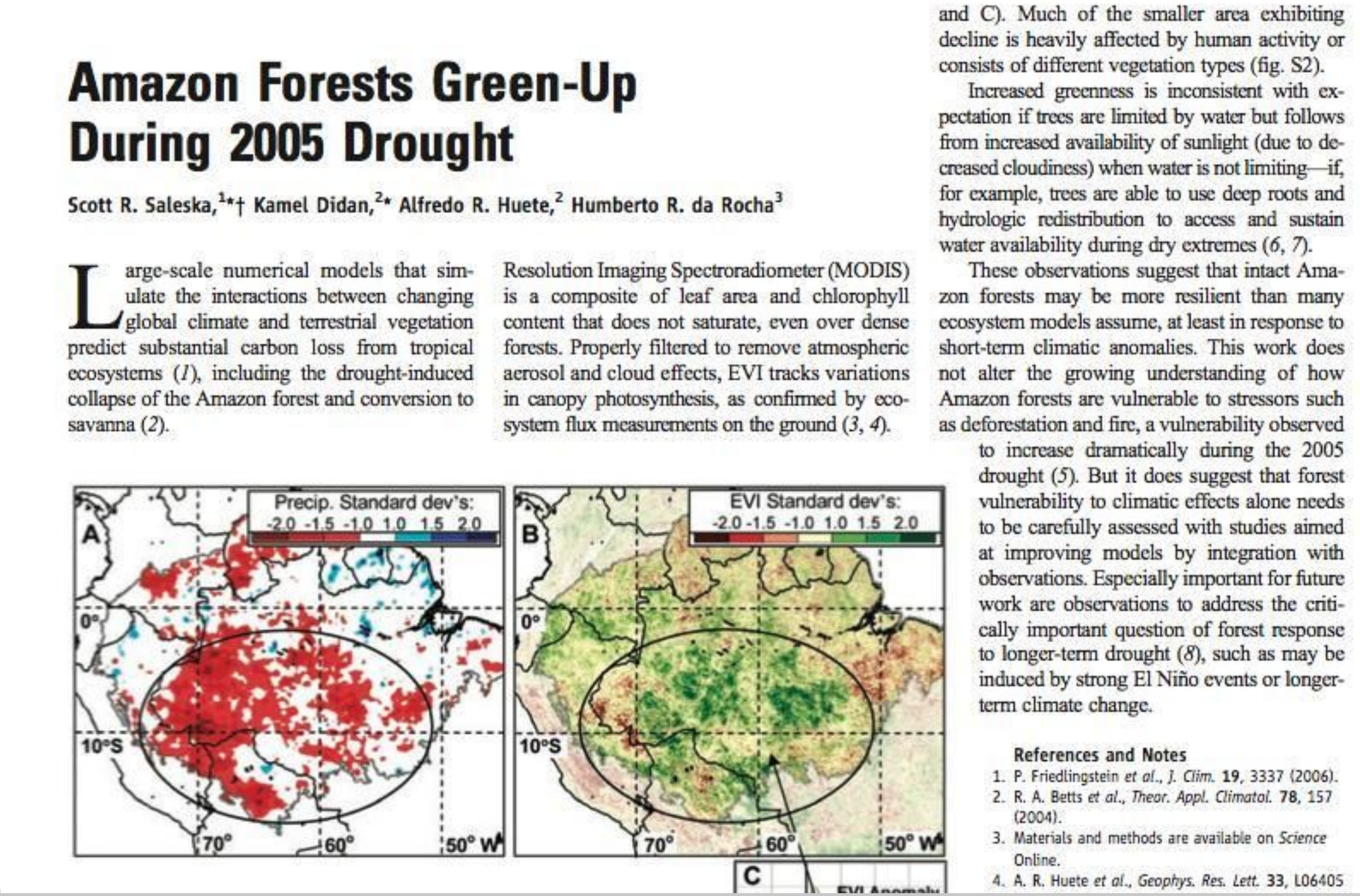
In this example we embed parts of the examples from the `kruskal.test` help page into a LaTeX document:

■■= data(airquality) library(ctest) kruskal.test(Ozone Month, data = airquality) which shows that the location parameter of the Ozone distribution varies significantly from month to month. Finally we include a boxplot of the data:

Source: http://www-cs-faculty.stanford.edu/~uno/lp.html

(..) *is a methodology that combines a programming language with a documentation language, thereby making programs more robust, more portable, more easily maintained* - Donald Knuth

Reproducible research



Amazon Forests Green-Up During 2005 Drought

Scott R. Saleska,[1]*† Kamel Didan,[2]* Alfredo R. Huete,[2] Humberto R. da Rocha[3]

Large-scale numerical models that simulate the interactions between changing global climate and terrestrial vegetation predict substantial carbon loss from tropical ecosystems (*1*), including the drought-induced collapse of the Amazon forest and conversion to savanna (*2*).

Resolution Imaging Spectroradiometer (MODIS) is a composite of leaf area and chlorophyll content that does not saturate, even over dense forests. Properly filtered to remove atmospheric aerosol and cloud effects, EVI tracks variations in canopy photosynthesis, as confirmed by ecosystem flux measurements on the ground (*3*, *4*).

and C). Much of the smaller area exhibiting decline is heavily affected by human activity or consists of different vegetation types (fig. S2).

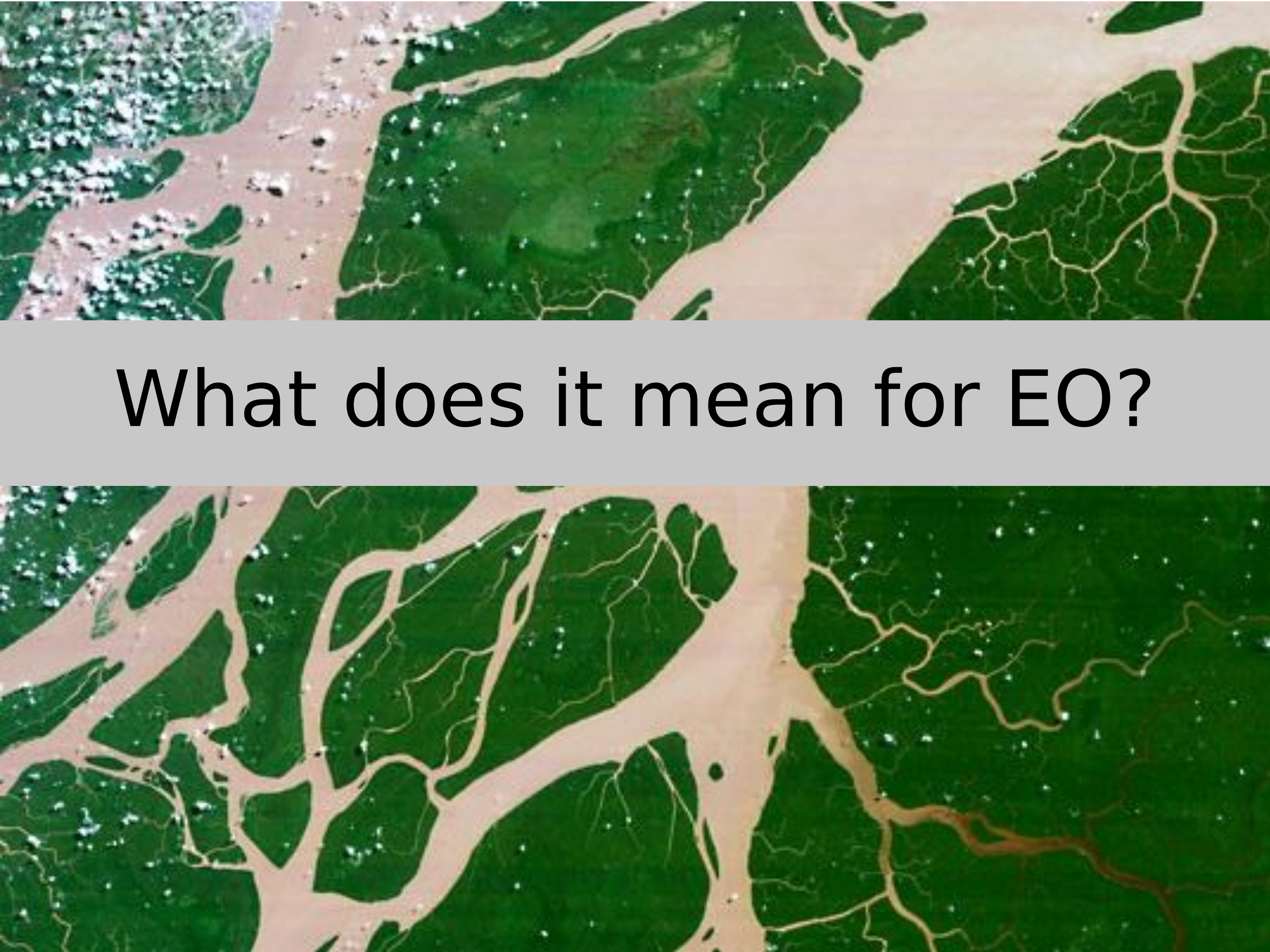
Increased greenness is inconsistent with expectation if trees are limited by water but follows from increased availability of sunlight (due to decreased cloudiness) when water is not limiting—if, for example, trees are able to use deep roots and hydrologic redistribution to access and sustain water availability during dry extremes (*6*, *7*).

These observations suggest that intact Amazon forests may be more resilient than many ecosystem models assume, at least in response to short-term climatic anomalies. This work does not alter the growing understanding of how Amazon forests are vulnerable to stressors such as deforestation and fire, a vulnerability observed to increase dramatically during the 2005 drought (*5*). But it does suggest that forest vulnerability to climatic effects alone needs to be carefully assessed with studies aimed at improving models by integration with observations. Especially important for future work are observations to address the critically important question of forest response to longer-term drought (*8*), such as may be induced by strong El Niño events or longer-term climate change.

References and Notes

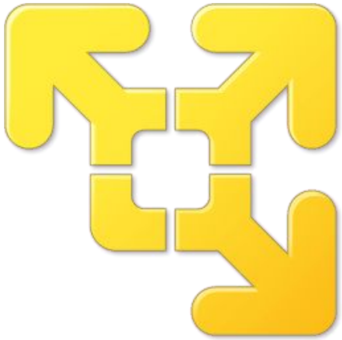
1. P. Friedlingstein *et al.*, *J. Clim.* **19**, 3337 (2006).
2. R. A. Betts *et al.*, *Theor. Appl. Climatol.* **78**, 157 (2004).
3. Materials and methods are available on *Science* Online.
4. A. R. Huete *et al.*, *Geophys. Res. Lett.* **33**, L06405

Science Magazine. 26 October 2007: Vol. 318 no. 5850 p. 612



What does it mean for EO?

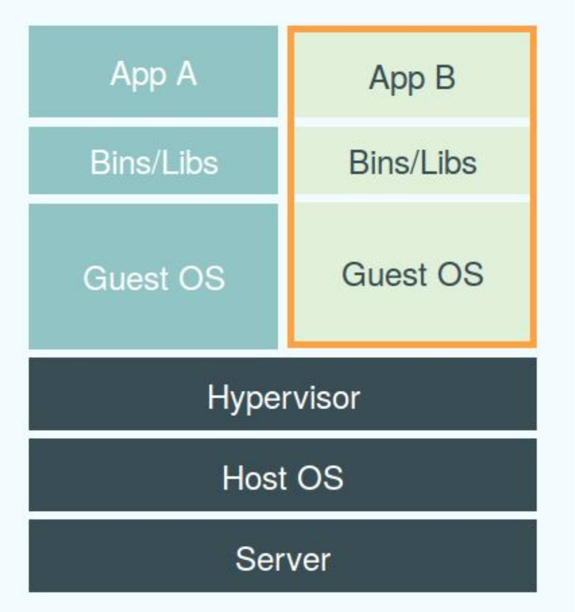
Virtualization



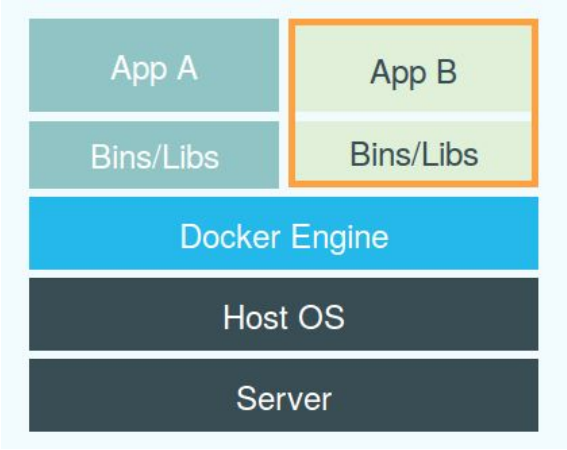
(...) the act of creating a virtual (rather than actual) version of something, including (but not limited to) a virtual computer hardware platform, operating system (OS), storage device, or computer network resources - Wikipedia

Docker

Virtual machine



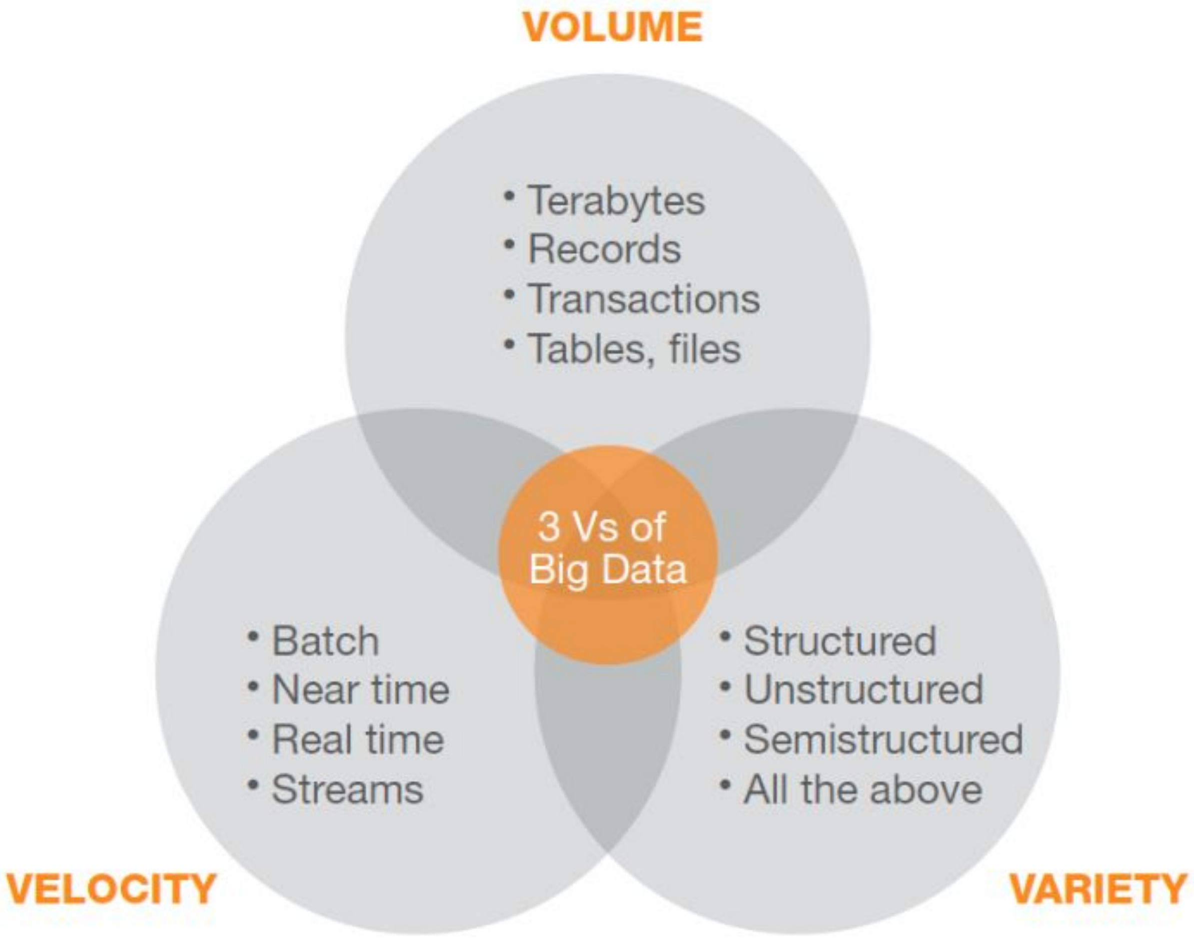
Docker container



Source: https://www.docker.com/whatisdocker/

Docker is an open-source project that automates the deployment of applications inside software containers - Wikipedia

Big data



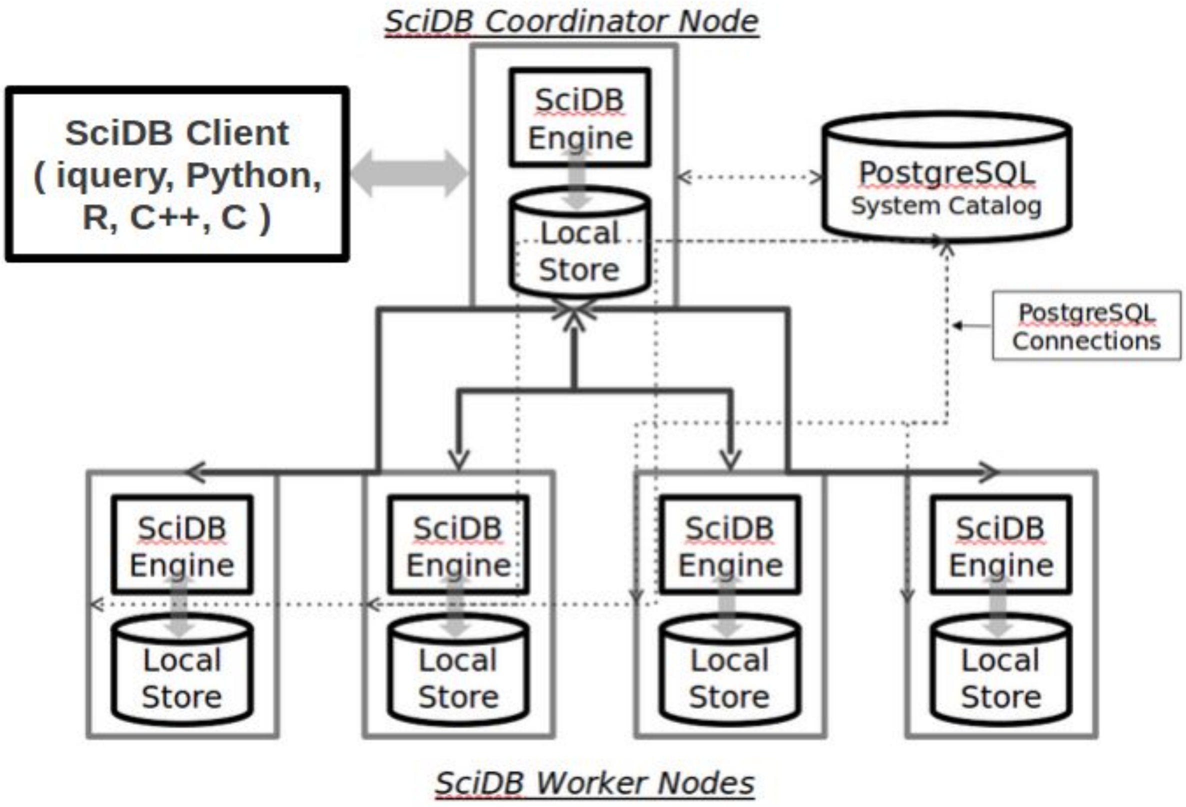
Source: http://www.digitalstrategy.ie/big-data-for-marketers/

~ When the sample is close to the Population ~

Array DBMS

(...) homogeneous collections of data items, sitting on a regular grid of one, two, or more dimensions. - Wikipedia

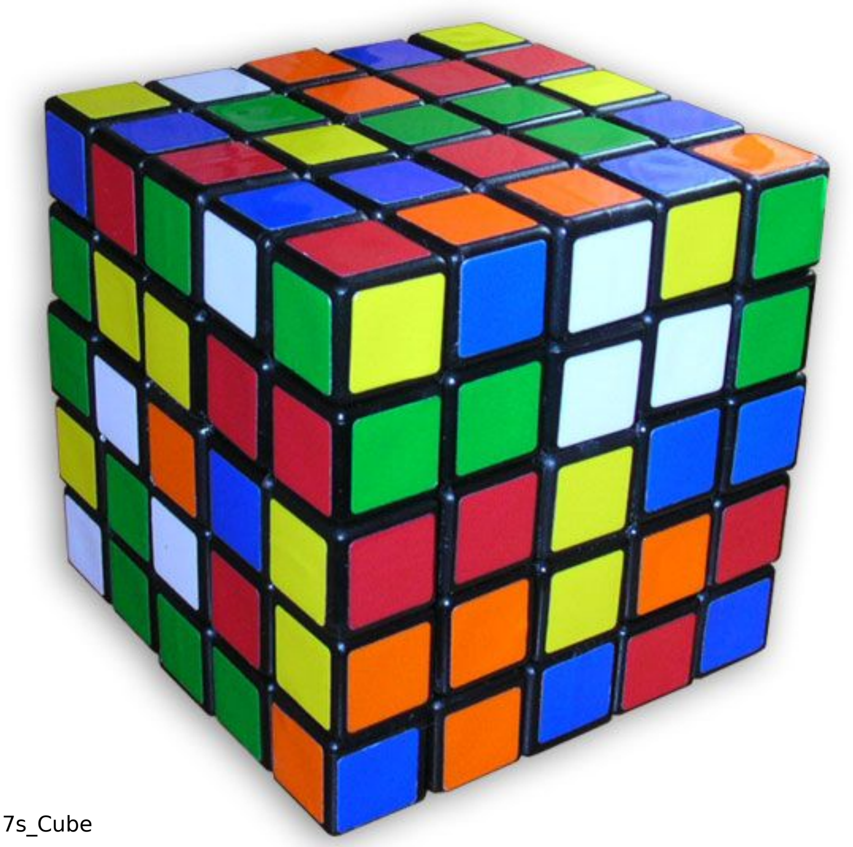
SciDB



Source: http://paradigm4.com/HTMLmanual/13.3/scidb_ug/ch01s02.html

It is an array database designed for multidimensional data management and analytics

SciDB Chunks



Source: http://en.Wikipedia.org/wiki/Professor%27s_Cube

Large arrays are split into chunks
which are distributed among instances

SciDB Array

```
CREATE ARRAY Simple_Array
< v1:   double,
v2 :    int64,
v3 :    string >
[I = 0:*, 5, 0,
J = 0:9, 5, 0];
```

Attributes
v1, v2, v3

Dim
I, J

Dim size
* is unbounded

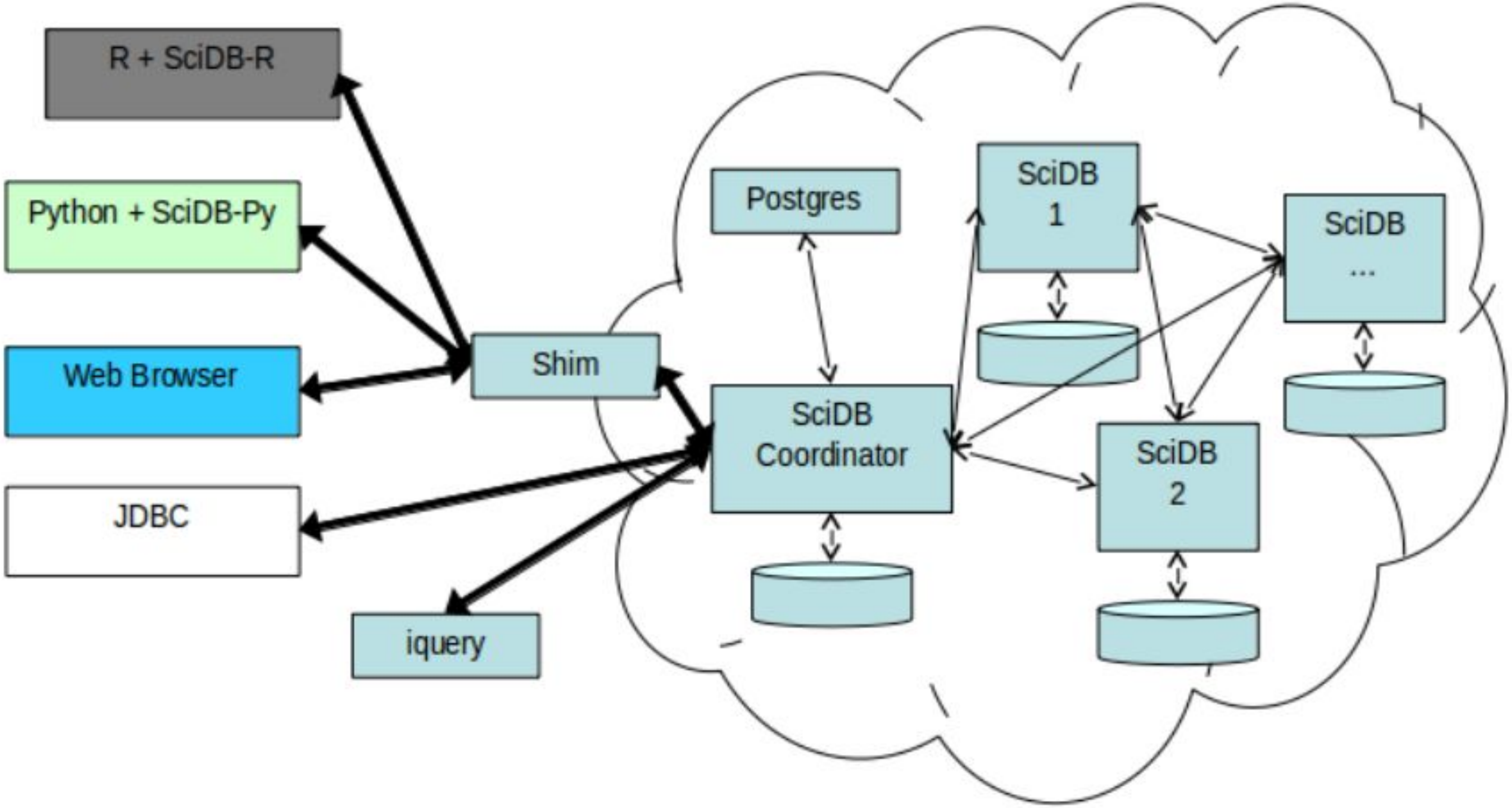
Chunk
size

Chunk
overlap

Source: http://paradigm4.com/HTMLmanual/13.3/scidb_ug/ch01s02.html

Attributes, dimensions and chunks.

SciDB Architecture

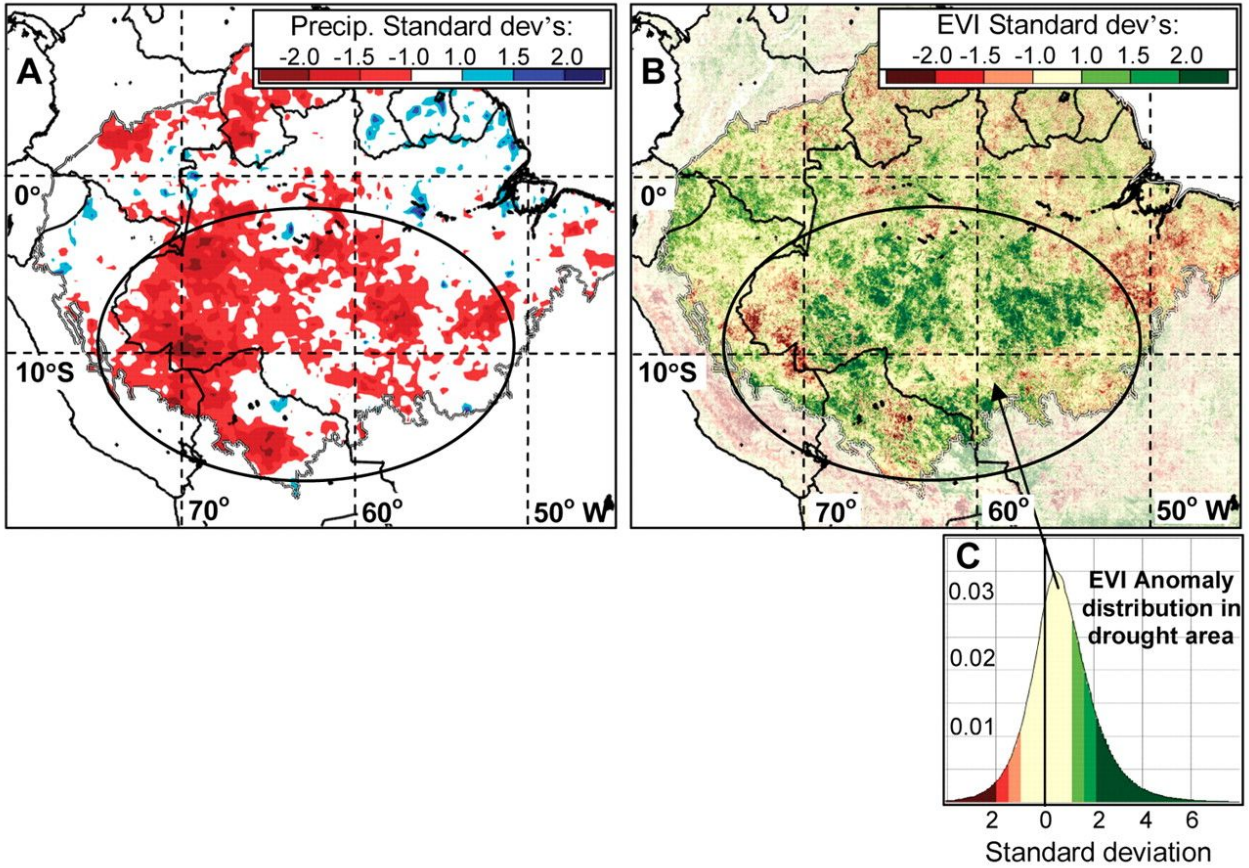


Source: http://paradigm4.com/HTMLmanual/14.8/scidb_ug/ch01s04.html

It uses a shared-nothing architecture

Reproducing an article

Did Amazon forests green up during 2005 drought?



Reproducing a paper using SciDB AFL

1. Load MOD09Q1 tiles
2. Extract pixels covering Amazon
3. Compute EVI2
4. Compute mean and STDEV for 2000-2006 and 2005
5. Join
6. Calculate anomalies

1 - MODIS Array

```
CREATE ARRAY MOD09Q1_BR_2000_2013
<red :       int16,
nir :        int16,
quality:     uint16>

[col_id = 48000:67199,    1014, 5,
row_id =  38400:52799,    1014, 5,
time_id = 0:9200,         1,   0];
```

Note the time dimension

2 - Extract pixels covering Amazon

```
store(
  between(
     filter(MOD09Q1_BR_2000_2013,

time_id % 46    >=   23 and
time_id % 46    <=   34 and
quality         =    4096

     ), 48000, 38400, 0, 67199, 52799, 321
  ), MODIS_AMZ_BQ_JAS
);
```

Selects the data using spatial (amazon),
temporal (JAS 2000-2006), and quality criteria

3 - Compute EVI2

```
store(
  apply(MODIS_AMZ_BQ_JAS,

evi2, 2.5*((nir-red)/(nir+2.4*red+(1*10000)))

  ),
  MODIS_AMZ_BQ_JAS_EVI2
);
```

Compute EVI for each cell

4 - Compute mean and STDEV

```
store(
  aggregate(
      filter(MODIS_AMZ_BQ_JAS_EVI2,
            time_id <= 229 or time_id >= 276
      ),
      avg(evi2)             as evi2_avg_jas_2000_2006,
      stdev (evi2)          as evi2_stdev_jas_2000_2006,
      col_id,
      row_id
  ),
  MODIS_AMZ_BQ_JAS_EVI2_AVG_2000_2006
);
```

The average of 2005 is calculated in a similar wsay and stored as **MODIS_AMZ_BQ_JAS_EVI2_AVG_2005**

5 - Join

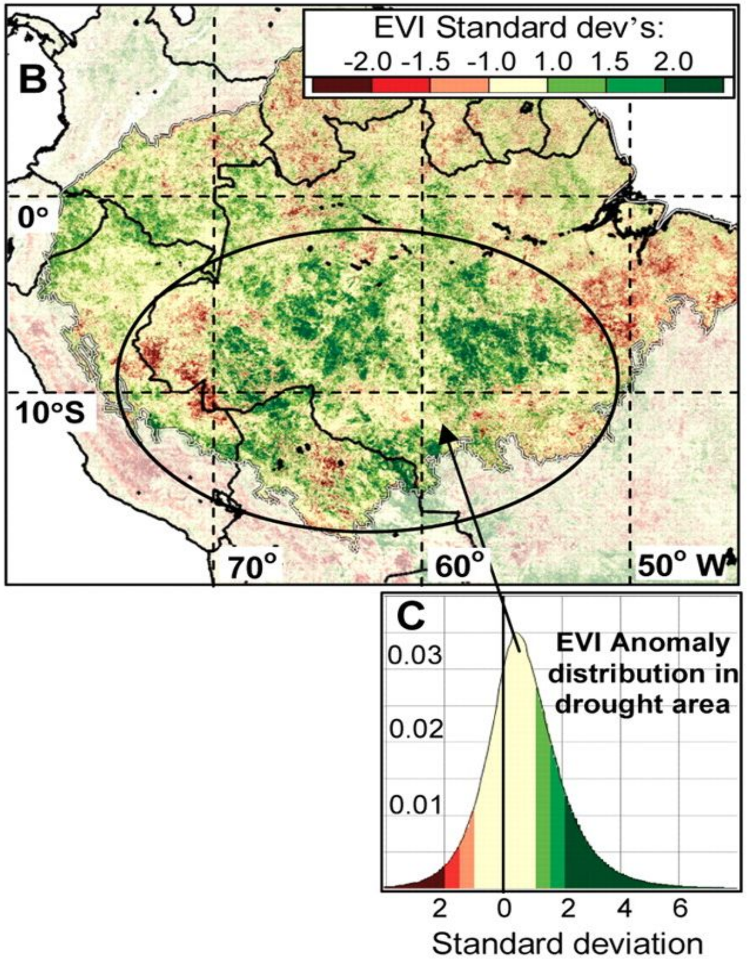
```
store(
  join(
     MODIS_AMZ_BQ_JAS_EVI2_AVG_2000_2006,
     MODIS_AMZ_BQ_JAS_EVI2_AVG_2005
  ),
  MODIS_AMZ_EVI2_COMP
);
```

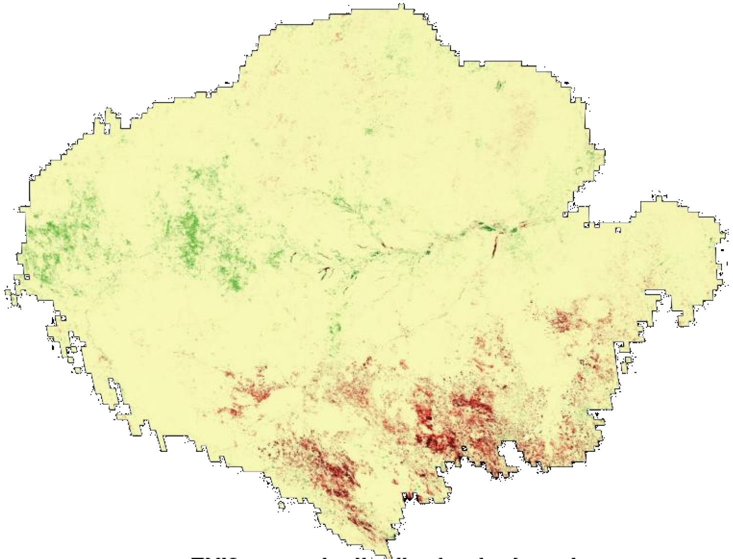
Join time series 2000-2006 to 2005

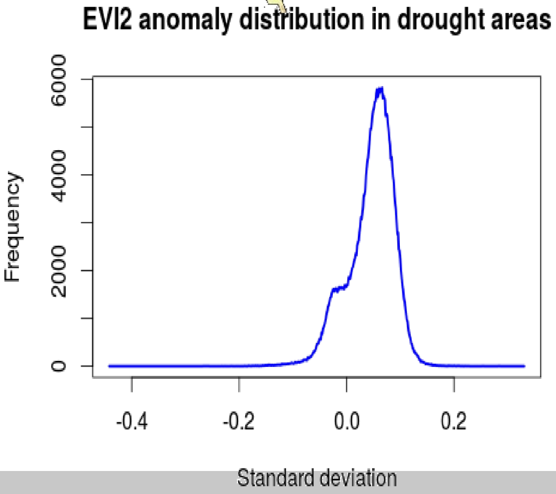
6 - Calculates Anomalies

```
store(
  apply(MODIS_AMZ_EVI2_COMP,
     evi_anomaly,

(evi2_avg_jas_2005 - evi2_avg_jas_2000_2006)
  / evi2_stdev_jas_2000_2006

  ),
  MODIS_AMZ_EVI2_ANOM
);
```

Results







https://github.com/albhasan/amazonGreenUp2005

Wrap up

- Reproducible EO research?
  - It is possible but it requires work
  - SciDB is not spatio-temporal enabled
- Literate programming
  - Array DB can manage large amounts of data. Perhaps through a central repository of known datasets (Landsat, MODIS...)
  - AFL is not the appropriate language

Resources

– Dockerized SciDB (install or compile)

https://github.com/albhasan/docker_scidb

– Reproduce an article running 2 scripts

https://github.com/albhasan/amazonGreenUp2005

Thanks!

In complex systems there is no relationship between information gathered and the decision made
Ed Horwood's Short Law #8