Analytics - BigData applied to Scientific and Spatial Development

Mauro Assis
assismauro@hotmail.com

EARTH SYSTEM SCIENCE CENTER

STRATEGIC GOAL

Development and improvement of earth systems models, monitoring networks and sociopolitics analisys, oriented to construction and analisys of climate changes scenarios and climate projection.

Demands I - Processes

- Reception, quality control, processing, storage and distribution of satellite data
- Monitoring of biomes land use changes
- Global and regional atmospheric model processing
- Ocean and ocean-atmosphere coupled models processing
- Field experiments, meteorologic images, oceanic imagens databases
- Run oceanic-atmospheric climate I

Demands II - Projects

- Research about nature in Brasil (climate, preservation, preservation economy)
- Investigation of climate changes and it’s future effects in the country
- Collaboration in global climate investigation efforts
- ...

IT infrastructure

- 20 petabytes of data
- 1 petabyte/year growing
- Cray XT-6 supercomputer
- Lots of workstations and computers available to the researchers

Hiring cloud services

- Flexible (on-demand)
- Cost-effective
- But…
- No legal framework
- Cultural barrier

Estudo de caso

How much does the Amazon weigh?

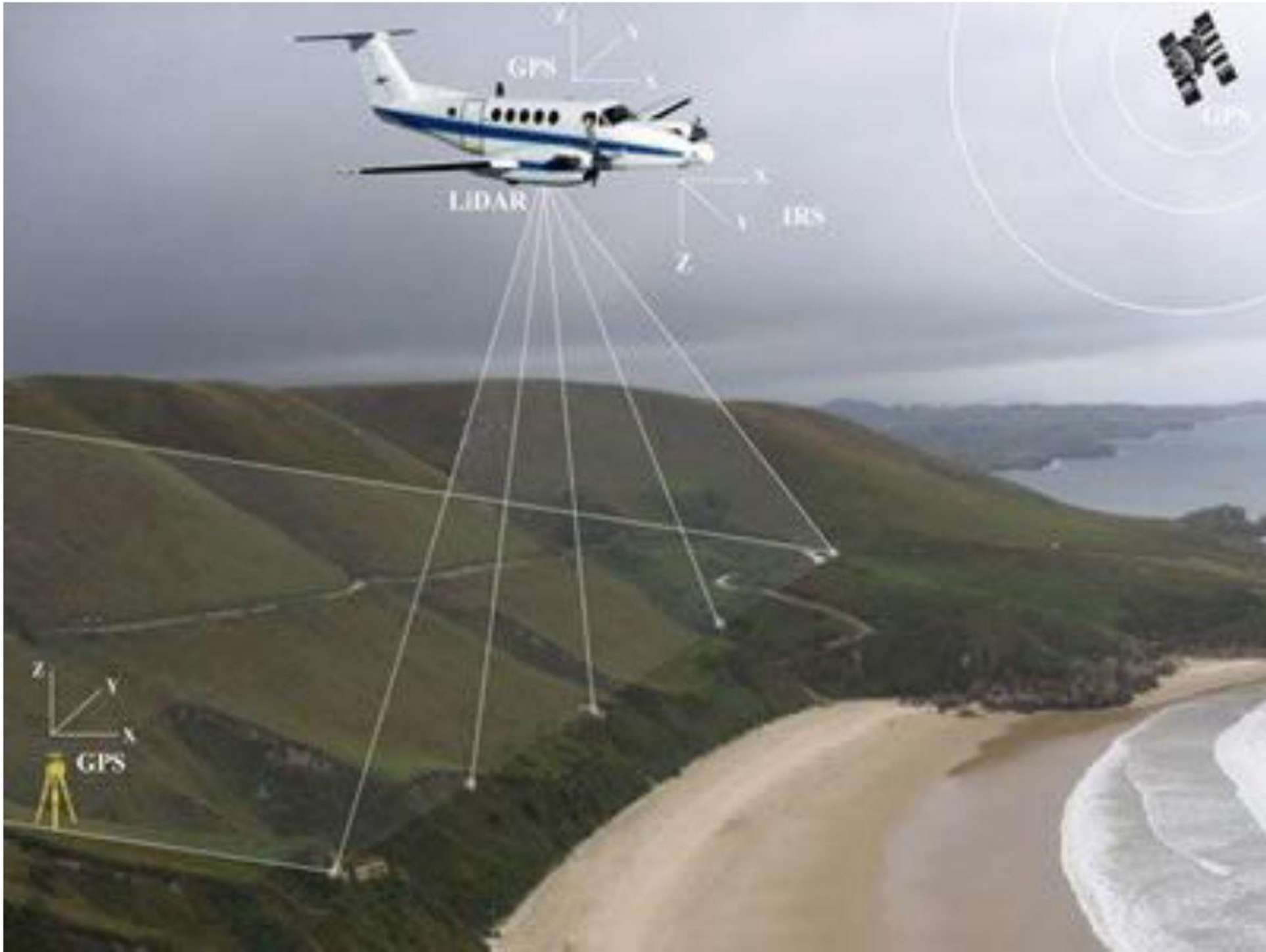
Previous question:

Why map the biomass of the Amazon forest?

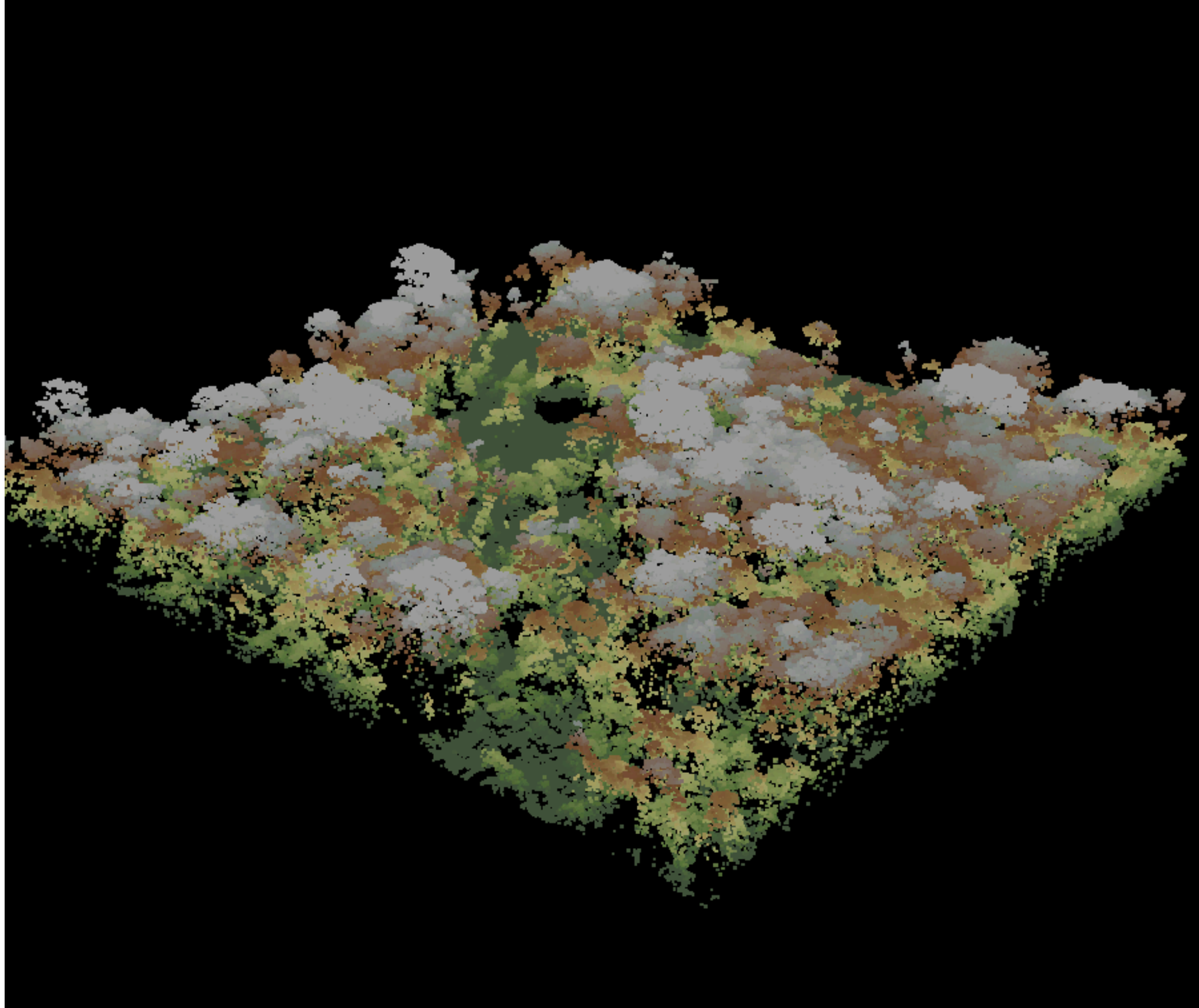
Biomass map process

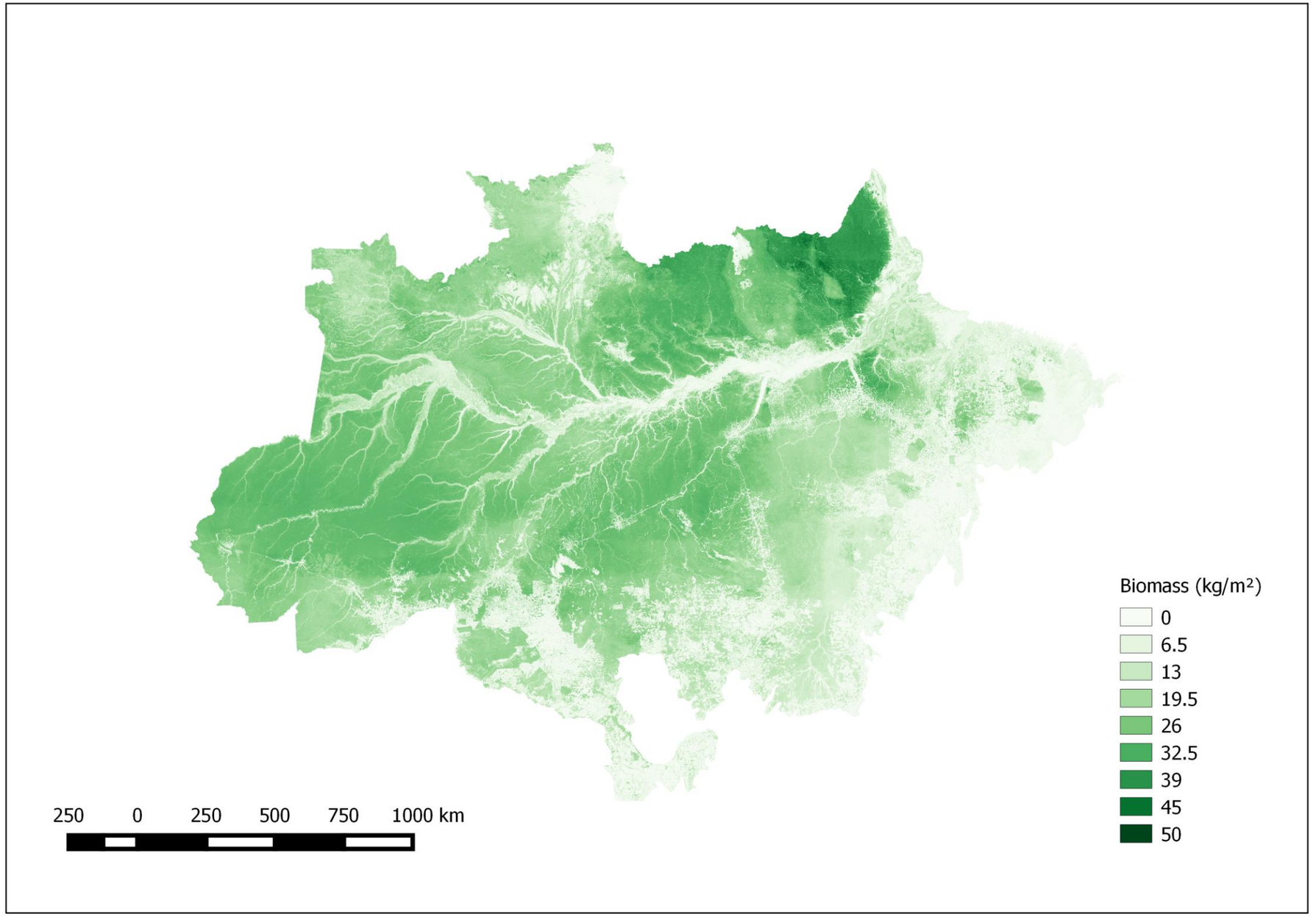
- 68 million pixels (250 x 250m)
- 4 million km² area
- ~1.000 LiDAR flights data
- Each flight: 6.5 billion of data recs
- 10 bands of satellite data for each pixel
- 4 to 6 h/map generation
- 16 CPU/32 gbyte RAM/21 Tb HD
- Random Forest algorithm
- Python H2O

LiDAR



LiDAR





Uncertainty map

- Propagate error from field to Random Forest extrapolation
- 1.000 biomass values normally distributed for each pixel
- A thousand maps to generate generate…
- … How??

La respuesta: Procesamiento analítico en AWS

- AWS contrata a un socio (DataRain)
- Dos PoCs
- Cuatro instancias de EC2 núcleos Linux 64 / 256 Gbytes cada una
- Ambiente Anaconda/H2O Python
- Script con mucho procesamiento paralelo
- Área amazónica dividida en 16 segmentos
- Dos operadores para ejecutar todo en 40 horas
- Descargamos los 1000 mapas y los resumimos en INPE
- Tardó unos 2 días en generar el mapa final

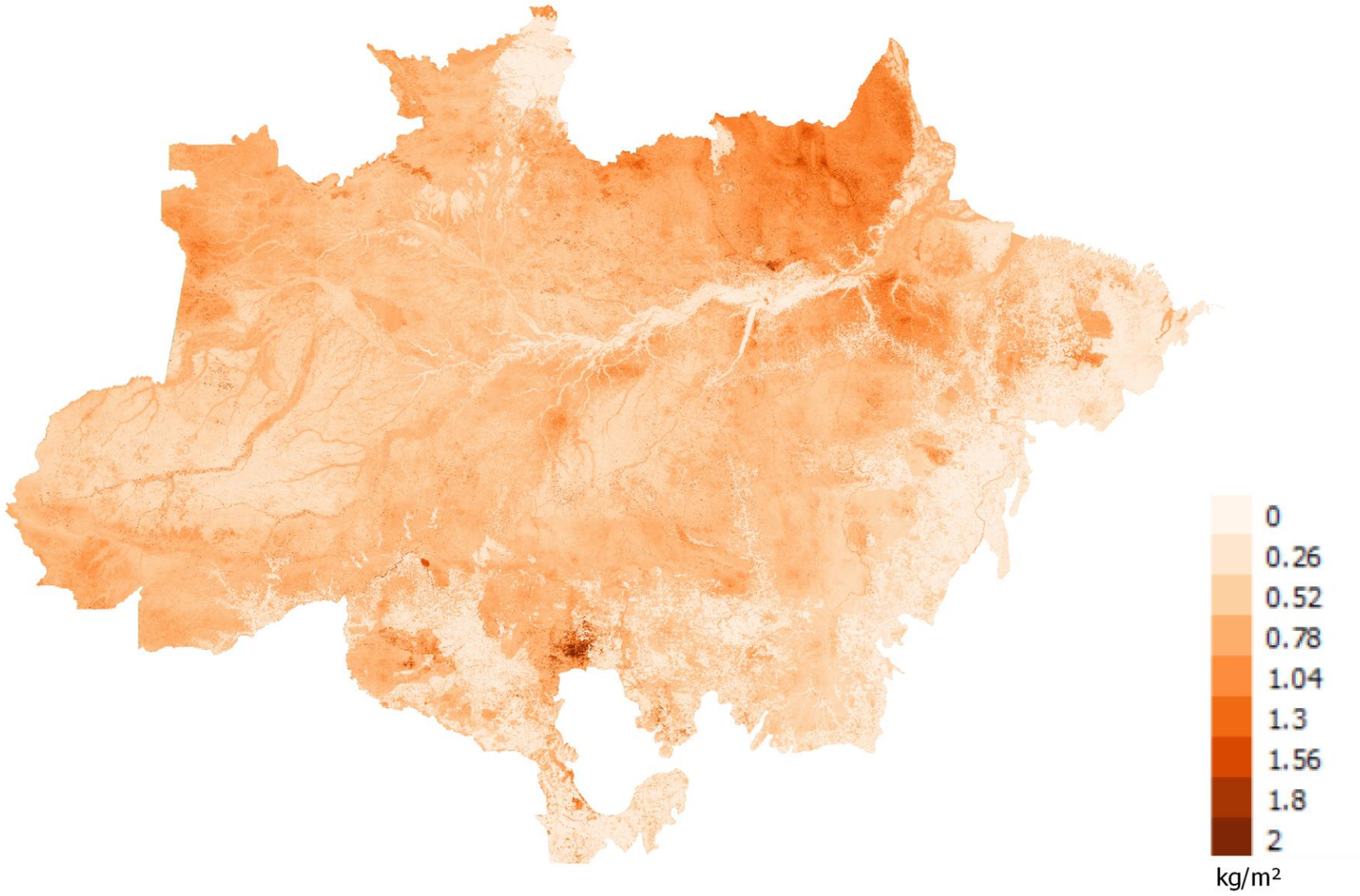
La respuesta: AWS Big Data Analytics

Four AWS Linux w/ 64 cores/256 GB RAM selected

- Anaconda/H2O Python environment setup
- EC2 AMI Replication (x4)
- Split Amazon area into 16 segments




Uncertainty Map



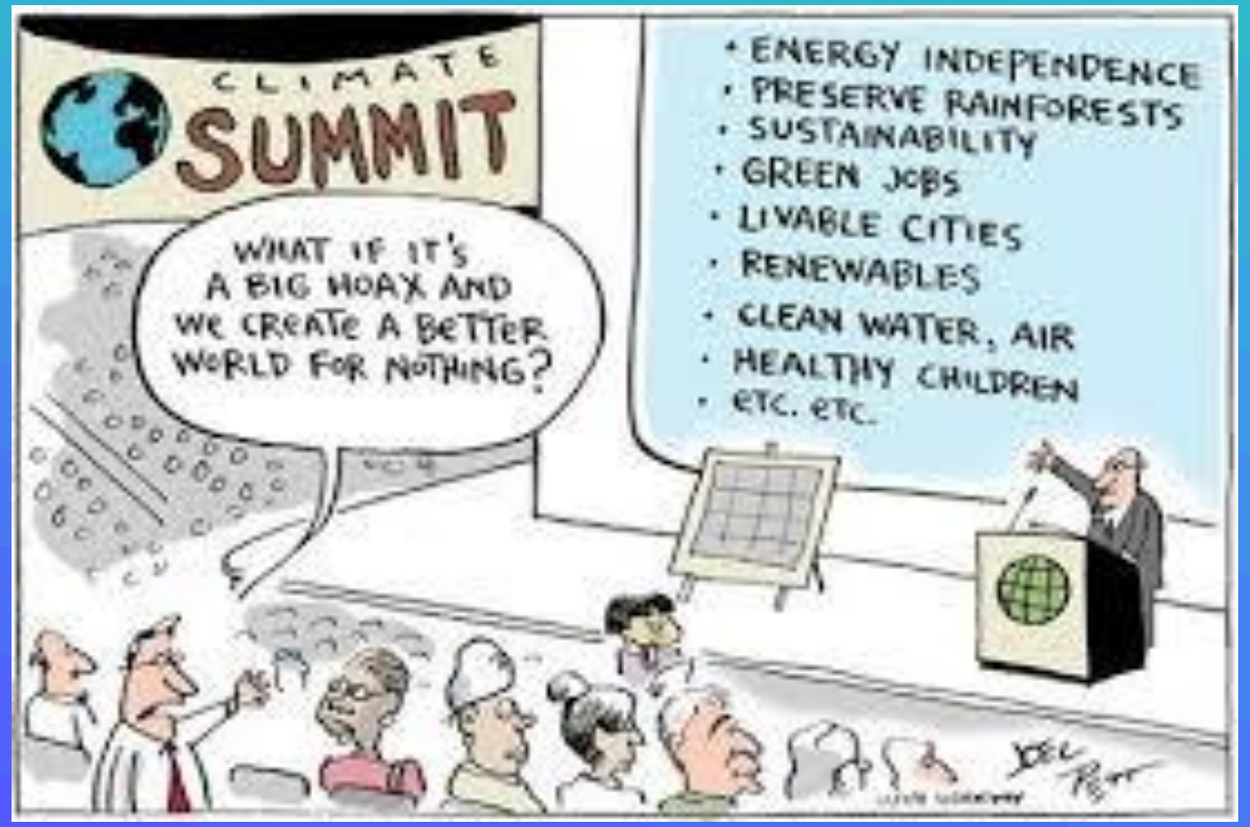
Main benefits

- Uncertainty map itself
- First time using cloud services at INPE
- 2.000 hours on-prem processing replaced by 160 hours on-cloud processing
- Final Map produced within project schedule
- High Return on investment (low cost, on-demand processing)

ROI

- LiDAR Fligts costs $3.000 each
- 1000 flights => $3M
- To update the model: 100~150 flights
- 150 flights => $450k
- Cost of map generation: $10.000
- Money saved in the next map updating:
- $3M – $450k – $10k = $2.54 M

¡Gracias!


CLIMATE
SUMMIT
WHAT IF IT'S A BIG HOAX AND WE CREATE A BETTER WORLD FOR NOTHING?
• ENERGY INDEPENDENCE
• PRESERVE RAINFORESTS
• SUSTAINABILITY
• GREEN JOBS
• LIVABLE CITIES
• RENEWABLES
• CLEAN WATER, AIR
• HEALTHY CHILDREN
• ETC. ETC.