Escuela de organización industrial

BIG DATA E INTELIGENCIA GEOESPACIAL: CASOS DE USO

JUAN PEDRO PÉREZ - SUNNTICS

01

Introducción
Información Geográfica y Big Data: Más allá del simple punto

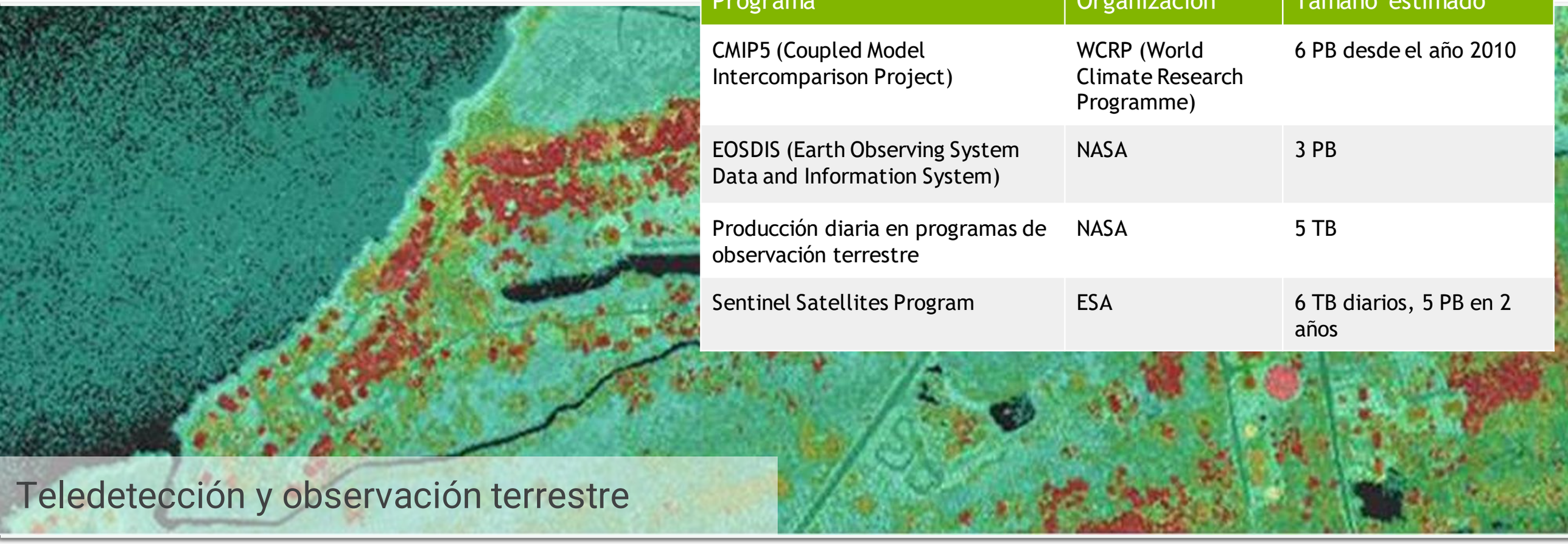
Algunas ideas clave…

Los datos geográficos siempre han sido insoportablemente “big”

No es lo mismo Big Data Geográfico que lo geo que hay en el Big Data

Por su naturaleza, es difícil crear metodologías estándar Big Data para los datos geográficos

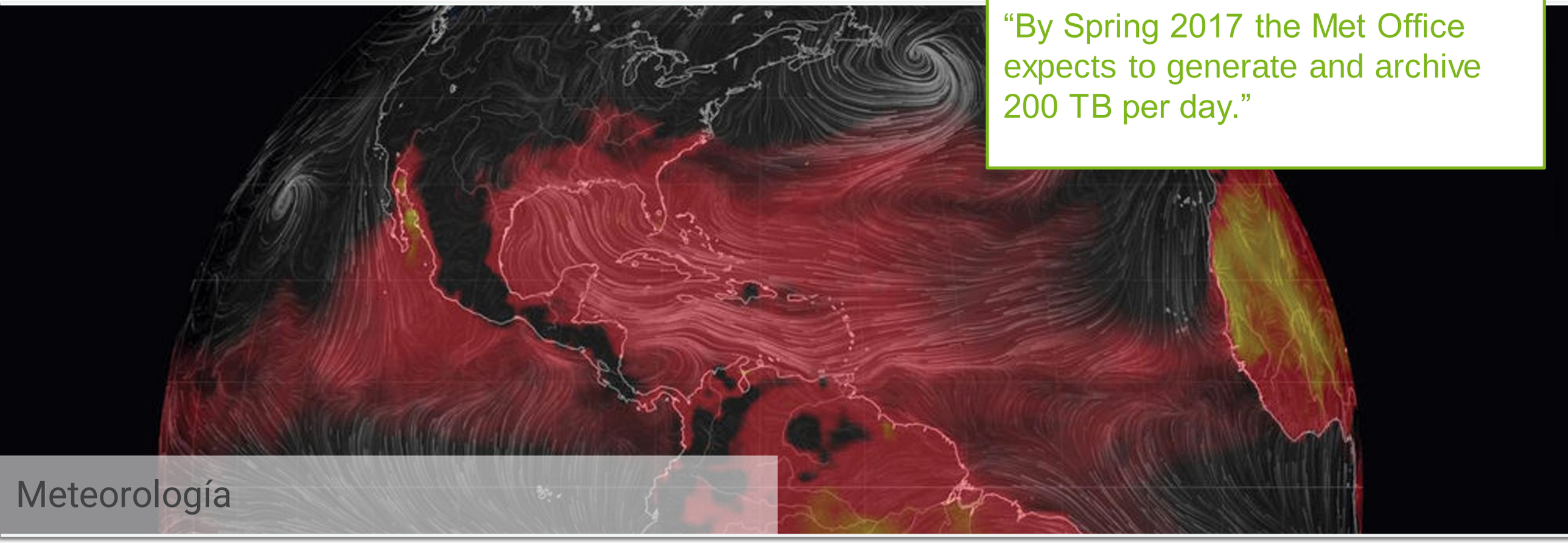
Big Data Geo



Programa	Organización	Tamaño estimado
CMIP5 (Coupled Model Intercomparison Project)	WCRP (World Climate Research Programme)	6 PB desde el año 2010
EOSDIS (Earth Observing System Data and Information System)	NASA	3 PB
Producción diaria en programas de observación terrestre	NASA	5 TB
Sentinel Satellites Program	ESA	6 TB diarios, 5 PB en 2 años

Teledetección y observación terrestre

Big Data Geo

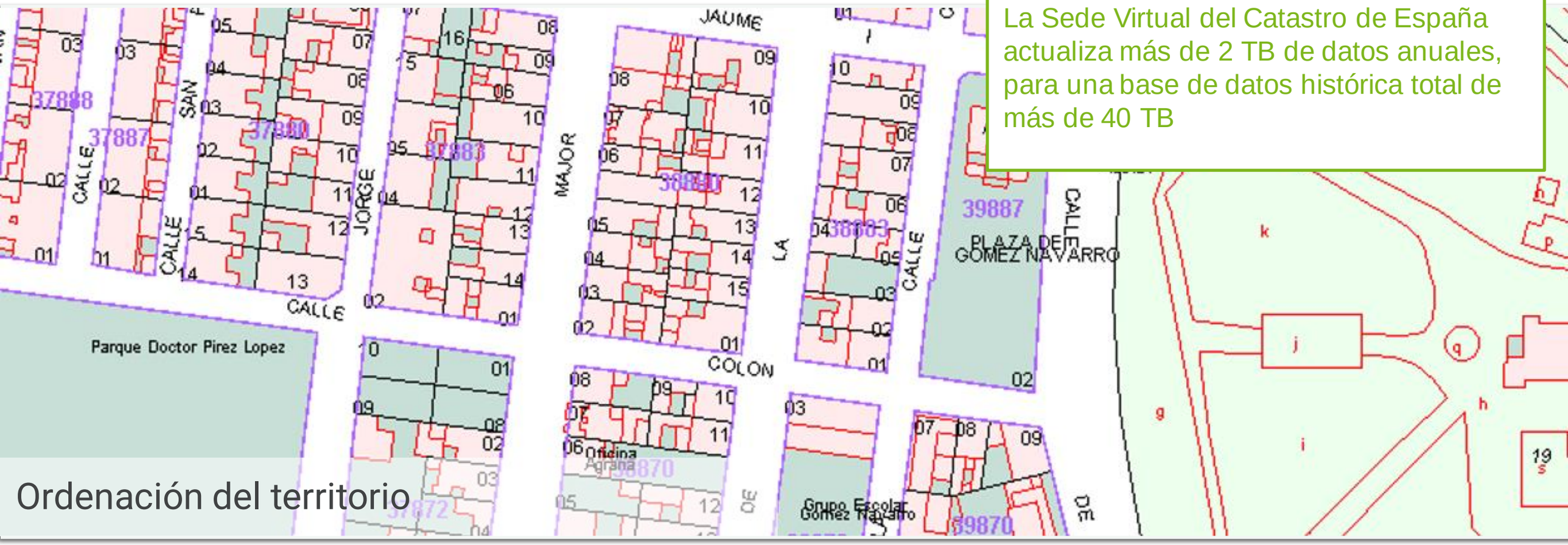


“By Spring 2017 the Met Office expects to generate and archive 200 TB per day.”

Meteorología

Big Data Geo

La Sede Virtual del Catastro de España actualiza más de 2 TB de datos anuales, para una base de datos histórica total de más de 40 TB



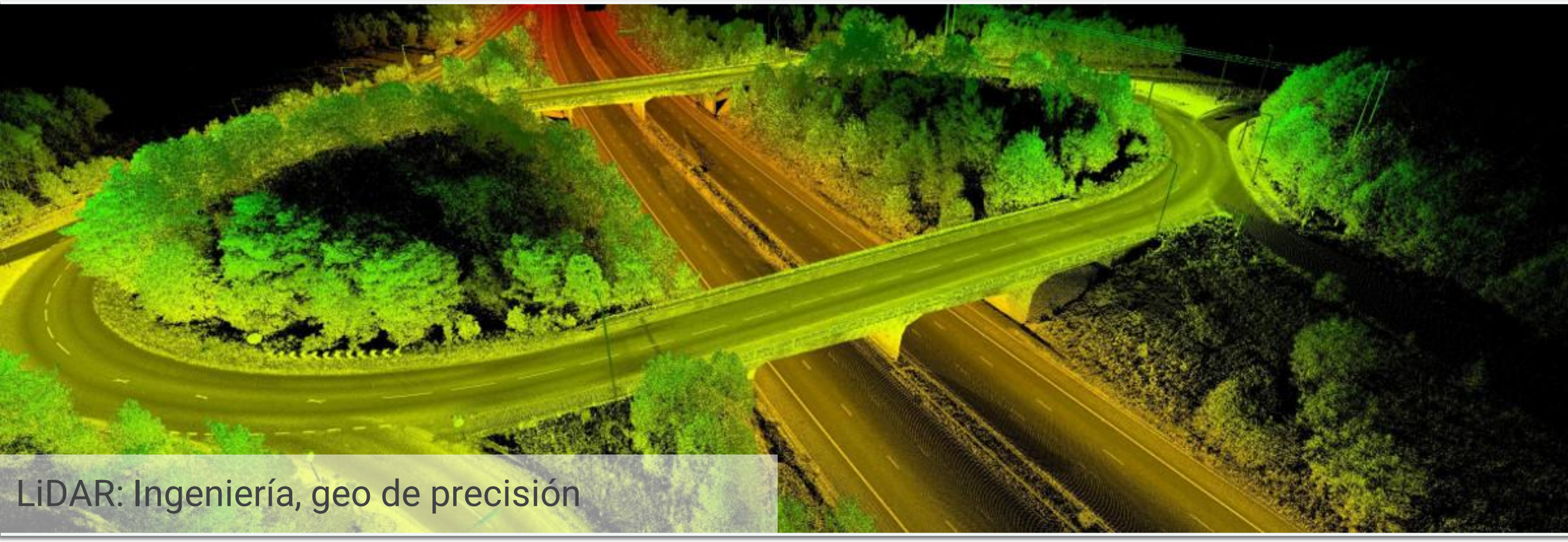
Ordenación del territorio

Big Data Geo

Google Maps: 20 PB
OpenStreetMaps: 800 GB

Cartografía de uso general, navegación

Big Data Geo



LiDAR: Ingeniería, geo de precisión

Evolución de los sensores embarcados en satélites

Año	Plataforma	Operador	Resolución (m / pixel)
1960	TIROS - 1	NASA	2500
1972	Landsat - 1	NASA	80
1984	Landsat - 5	NASA / NOAA	30
2001	QuickBird	Digital Globe	2.62 / 0.65 (Pan)
2015	Sentinel - 2	ESA	10 / 60
2016	WorldView - 4	Digital Globe	0.31
2020	Landsat - 9	NASA / USGS	30 / 15 (Pan)

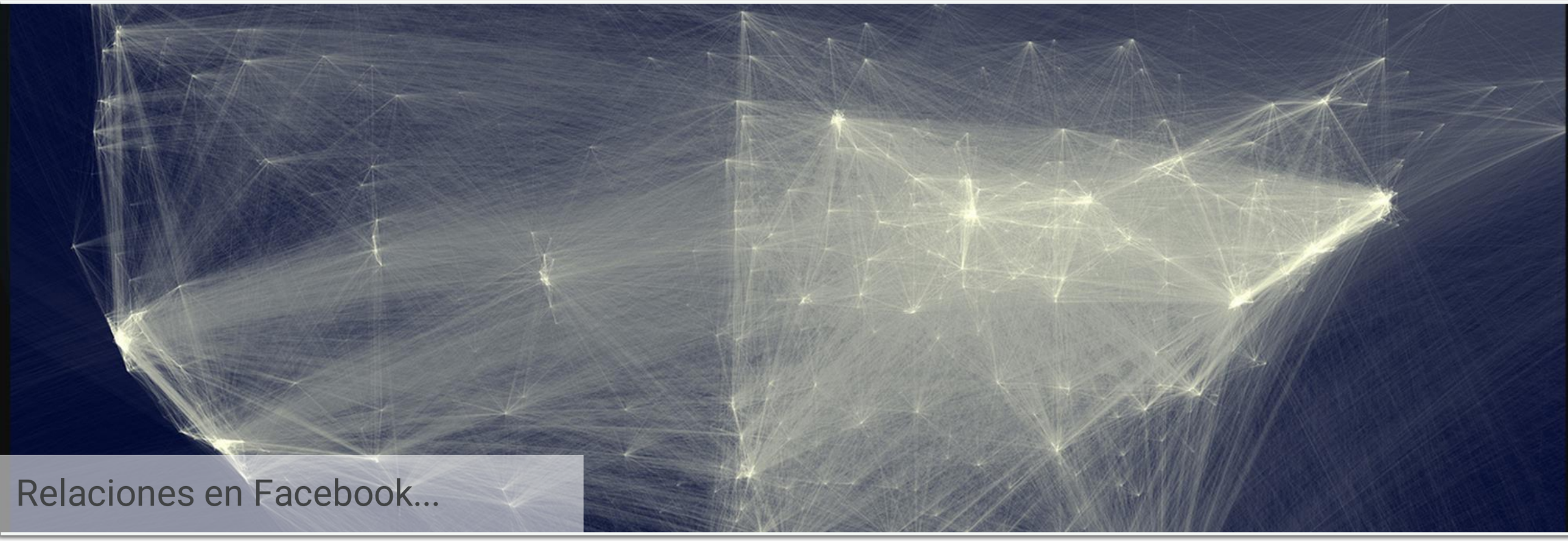
Algunas ideas clave…

Los datos geográficos siempre han sido insoportablemente “big”

No es lo mismo Big Data Geográfico que lo geo que hay en el Big Data

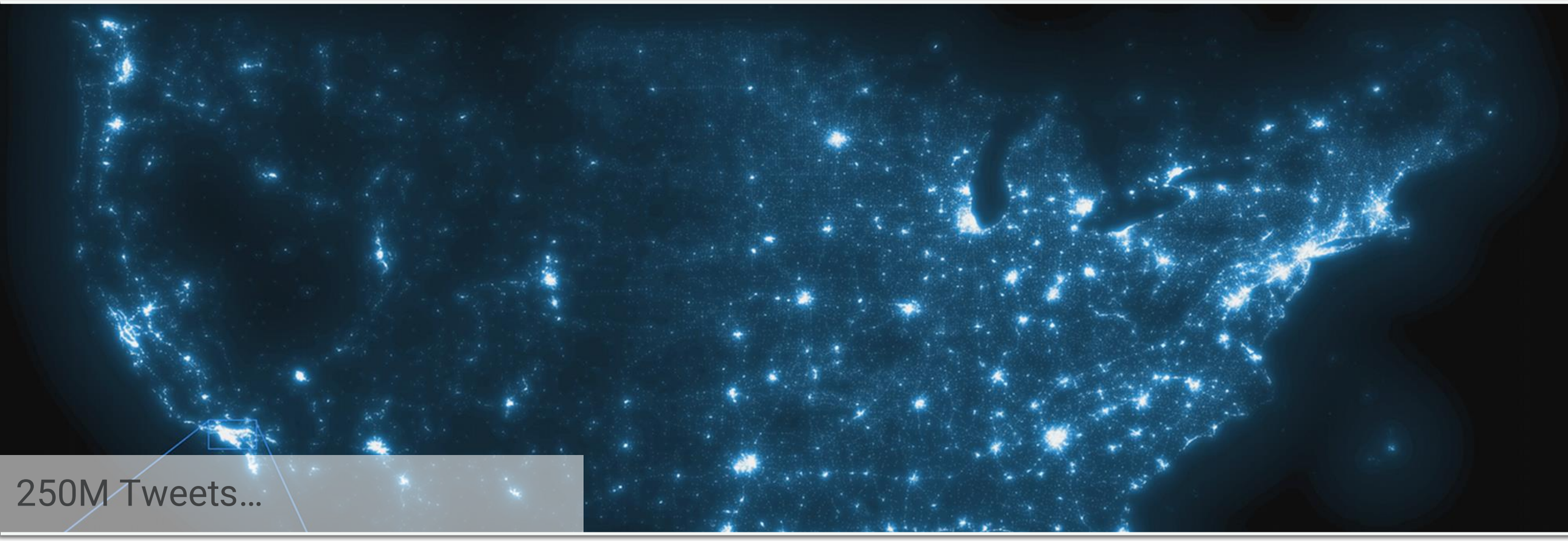
Por su naturaleza, es difícil crear metodologías estándar Big Data para los datos geográficos

Visualizando lo geo del Big Data…

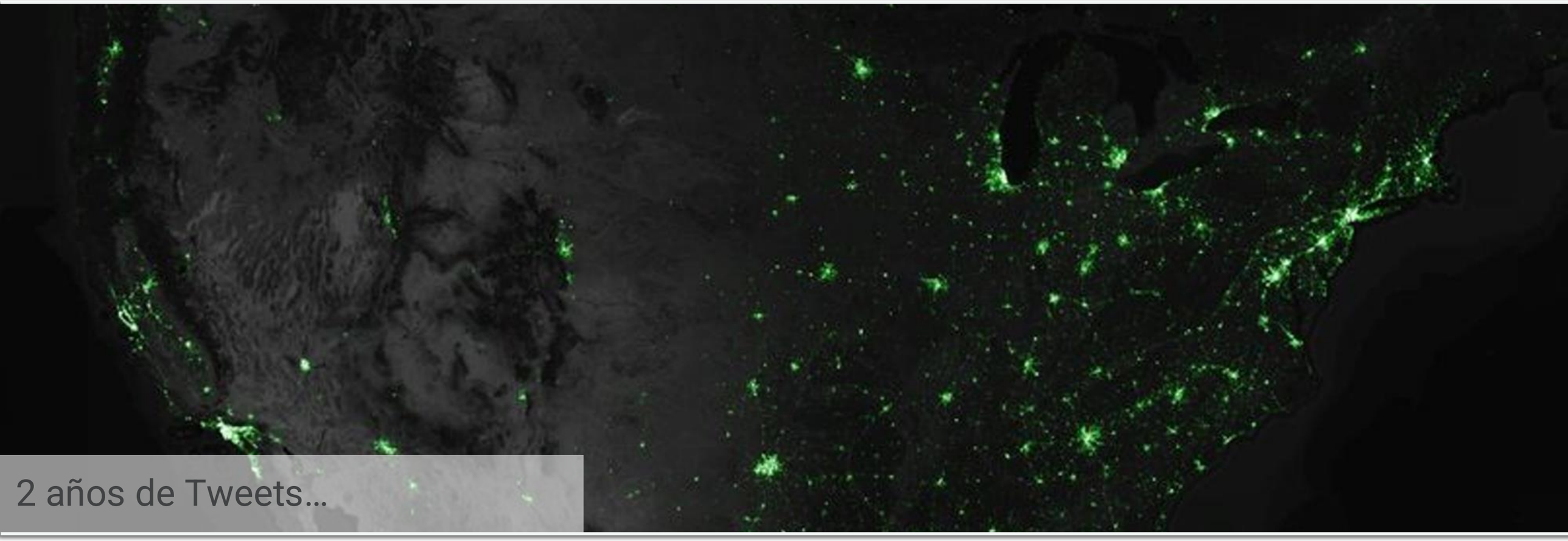


Relaciones en Facebook…

Visualizando lo geo del Big Data…

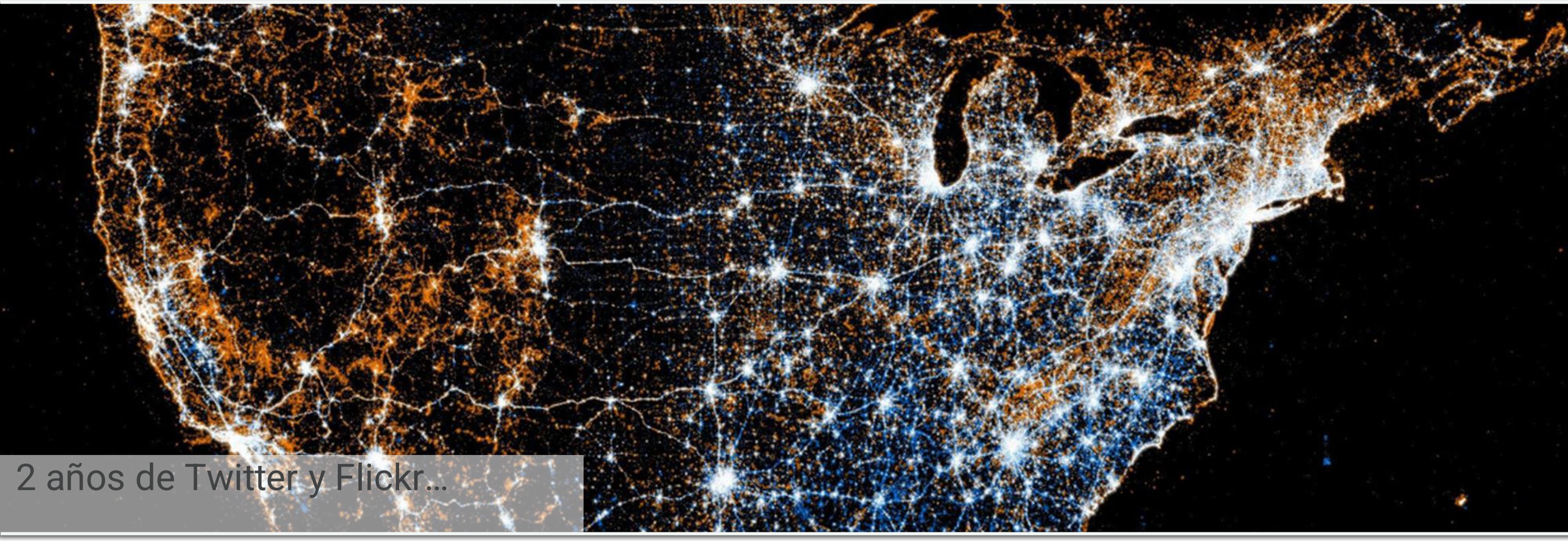


Visualizando lo geo del Big Data…

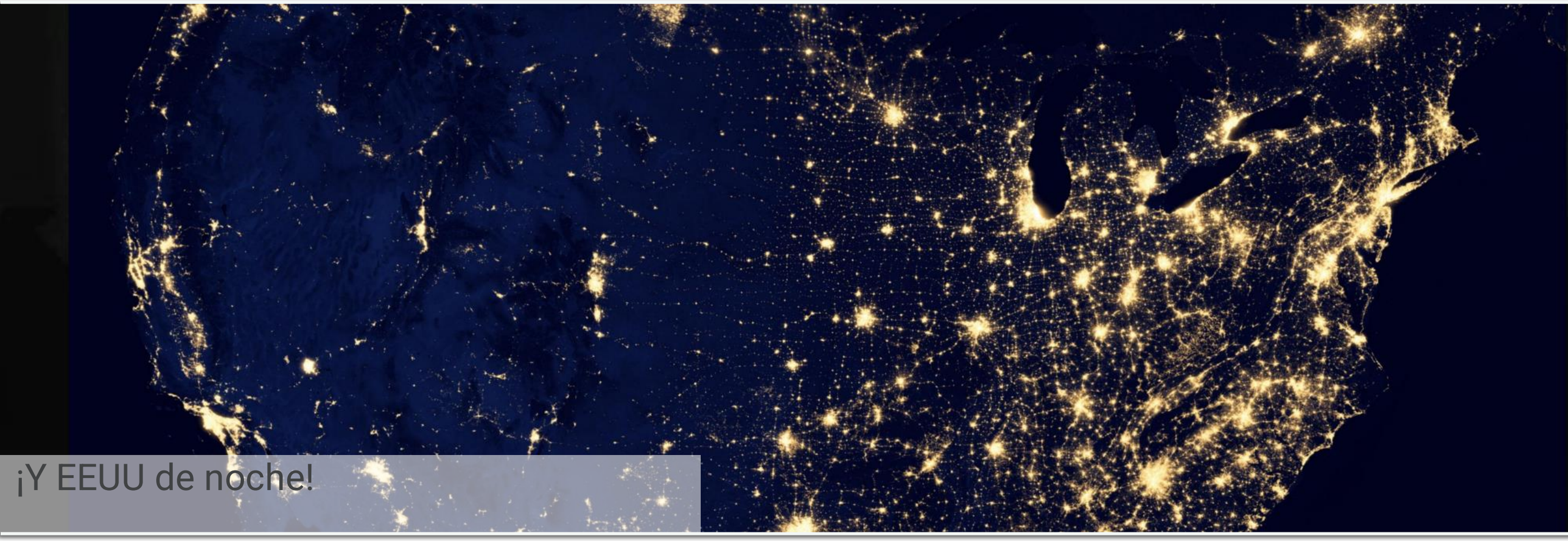


2 años de Tweets…

Visualizando lo geo del Big Data…

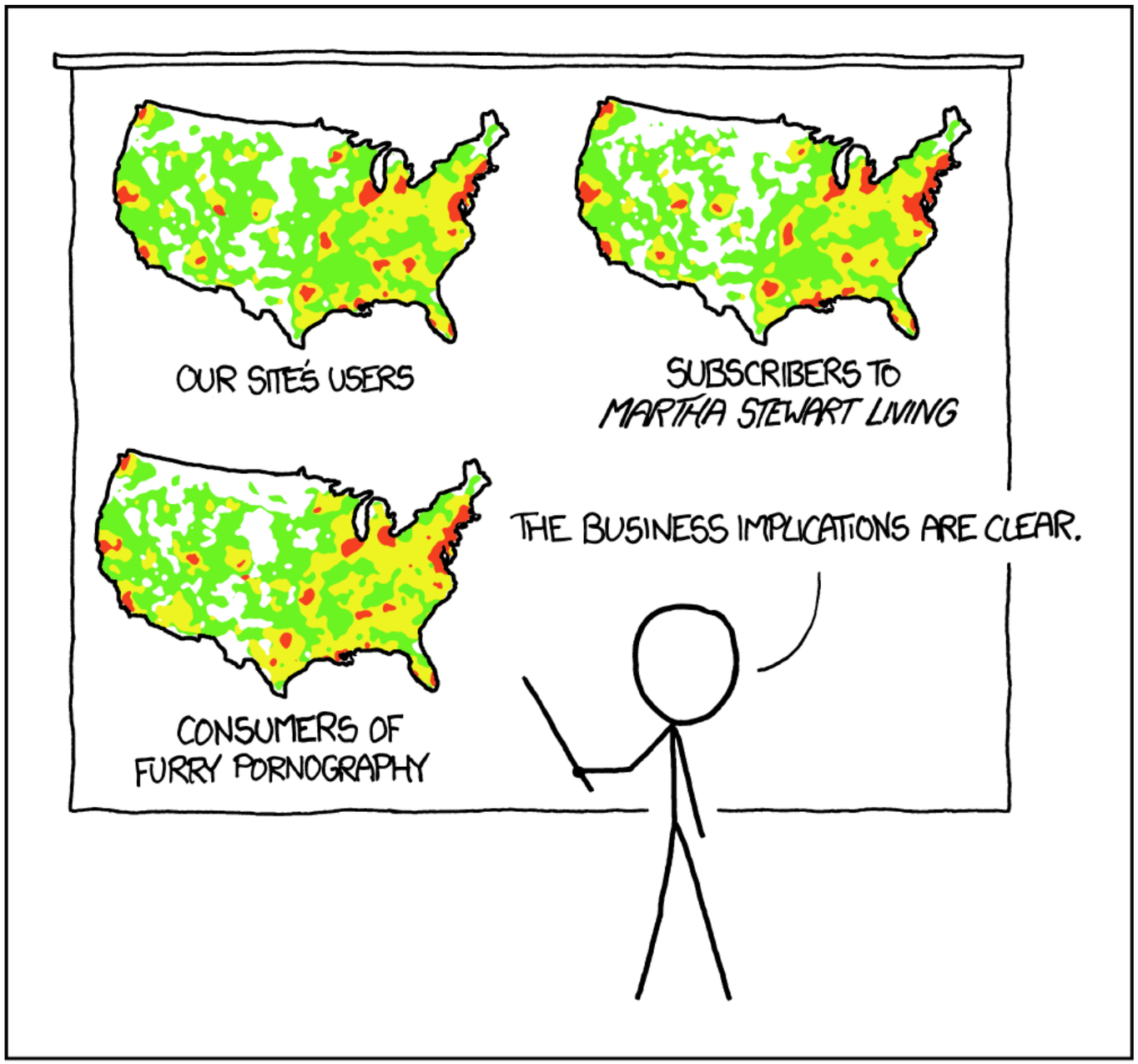


Visualizando lo geo del Big Data…



¡Y EEUU de noche!

Inferencia espacial…



PET PEEVE #208:
GEOGRAPHIC PROFILE MAPS WHICH ARE BASICALLY JUST POPULATION MAPS

Big Data Geo vs. Lo Geo del Big Data

Generalmente, lo Geo del Big Data se basa en el geoposicionamiento puntual de un conjunto de datos

El Big Data Geo va más allá, explorando las relaciones topológicas entre puntos, líneas y polígonos

Podríamos decir que lo Geo del Big Data es una importantísima y rica fuente de datos para el Big Data Geo

Algunas ideas clave…

Los datos geográficos siempre han sido insoportablemente “big”

No es lo mismo Big Data Geográfico que lo geo que hay en el Big Data

Por su naturaleza, es difícil crear metodologías estándar Big Data para los datos geográficos

Ciencia de datos geográficos

El científico de datos geográficos:

- es un ejemplo especializado de científico de datos
- tiene en cuenta la naturaleza dual de la información geográfica (IG)
- modeliza los problemas sobre una base espacial, utilizando diversos modelos de datos de IG
- aprovecha al máximo las relaciones topológicas que existen en la IG
- el científico de datos geográficos es un científico de datos, y como tal, procesa datos. El objetivo no es hacer mapas per se (eso es un cartógrafo, habilidades que también tiene que tener un buen científico de datos espaciales)

Ciencia de datos geográficos

El científico de datos geográficos se pregunta constantemente:

- ¿Qué hay en un lugar? ¿Qué impacto territorial tienen las actividades que nos interesan? ¿Qué posibilidades hay que ubicar sobre el territorio un conjunto de datos preexistentes?

- ¿Qué extensión y forma tienen estos elementos territoriales? ¿Qué modelo de datos se ajusta mejor a la resolución del problema entre manos?

- ¿Qué propiedades temáticas tienen nuestros datos o fenómenos territoriales?

- ¿Cómo se relacionan espacialmente estos fenómenos con el resto?

- ¿Qué metodologías de análisis podemos aplicar para extraer conclusiones del análisis combinado de las propiedades espaciales y temáticas de estos fenómenos?

La IG tiene una naturaleza dual…

La IG tiene una doble dimensión:

- **geométrica:** describe la ubicación, forma, extensión, etc. de un datos espacial
- **temática:** describe las propiedades temáticas de un elemento espacial

Relaciones en la IG

En la IG, se darán relaciones entre datos a dos niveles:

- **relaciones alfanuméricas:** las que se dan en la dimensión temática o alfanumérica de la IG. Son análogas a las que se dan, por ejemplo, en una base de datos relacional convencional. Son explicitas

- **relaciones topológicas:** exclusivas de la IG, se dan en la dimensión geométrica de la IG, y emanan, intrínseca e implicitamente, de la forma, ubicación, tamaño y posición relativa de unos datos espaciales con respecto al resto

Modelos de IG

Existen muchos, pero básicamente son variaciones de dos modelos fundamentales:

- **vectorial:** la dimensión geométrica de la IG se basa en modelizar elementos territoriales con puntos, líneas y polígonos, mientras que la dimensión temática va a una estructura tabular aneja

- **ráster:** la dimensión geométrica de la IG se basa en una rejilla regular de recubrimiento completo de un área, mientras que la dimensión temática se almacena en cubos temáticos asociados a cada celda del ráster

Big Data Geo

El procesamiento de la información geográfica es “big” debido a:

- el volumen inherente de los juegos de información geográfica
- lo intenso del procesamiento topológico al que se ve sometida

Ambos factores llevan inevitablemente a estrategias de computación distribuida

Problemas del Big Data Geo

Existen algunas dificultades para llegar a crear procedimientos generalistas en procesamiento distribuido de IG:

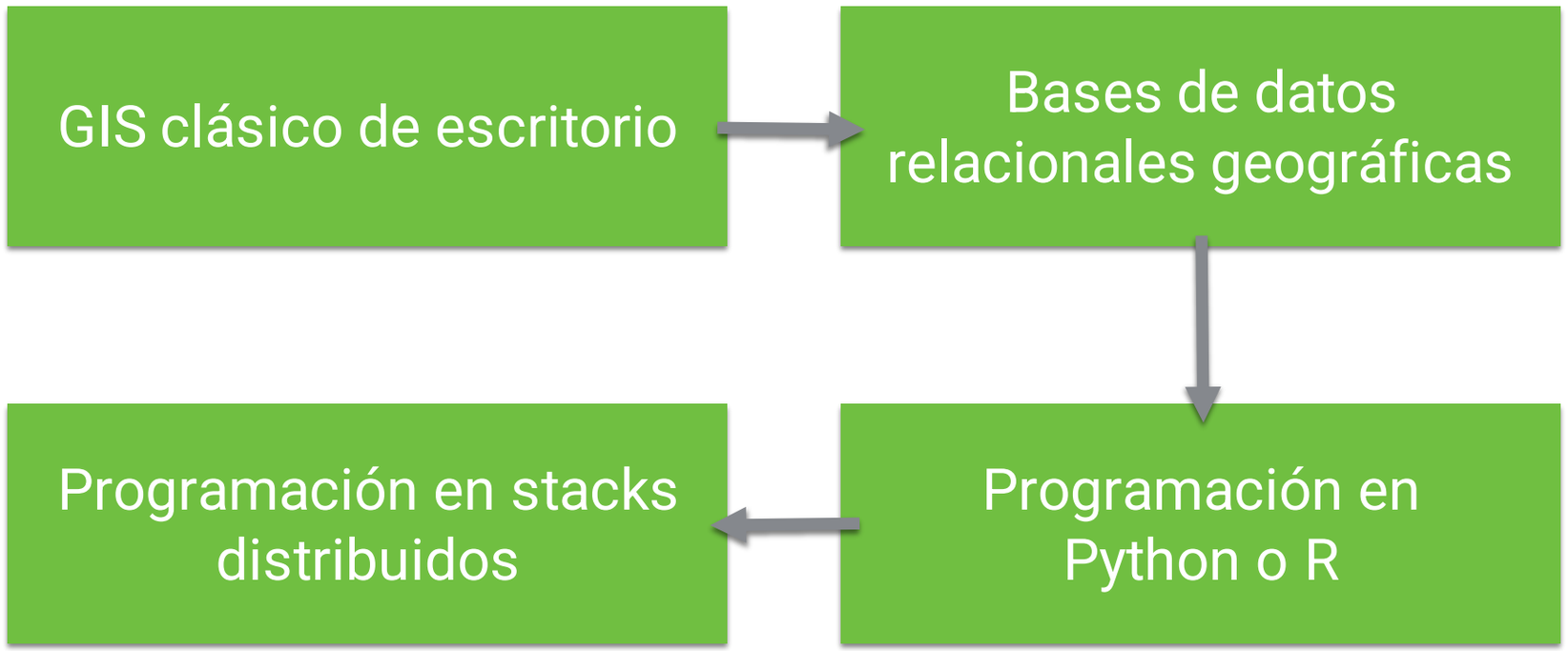
- la IG está implícita y fuertemente relacionada entre sí por medio de las relaciones topológicas
- los algoritmos tradicionales suelen ser por tanto monolíticos, precisando del acceso constante al total de los datos
- tradicionalmente, escalábamos hacia arriba
- hay que reescribir constantemente variaciones personalizadas de algoritmos clásicos para el problema en cuestión

Práctica con CARTO

02

Herramientas & Stacks para la Ciencia de Datos Geográfica

Herramientas



GIS clásico de escritorio

- Son programas de escritorio monolíticos para tratar y analizar IG
- Existen desde los años 70 y sus procedimientos conforman el corpus analítico clásico de esta ciencia (GIScience)
- Suele ser el primer paso en el entrenamiento de los científicos de datos espaciales
- Ejemplos: QGIS (software libre), GRASS (software libre), ArcGIS (propietario)

Bases de datos geográficas

- Son bases de datos relacionales de nivel corporativo que pueden almacenar geometrías y que tienen, además del clásico motor relacional, un motor topológico para rastrear relaciones topológicas

- Los datos se tratan con SQL, utilizando una variante espacial para rastrear relaciones topológicas

- Todos los algoritmos de la GIScience pueden reproducirse sin cambios sobre grandes bancos de datos (escalabilidad vertical)

- Alta expresividad para crear nuevos algoritmos

- Poseen todas las ventajas (e inconvenientes) de las BDR, como la multiconcurrencia

- Forman el núcleo de las Infraestructuras de Datos Espaciales (IDE)

- Ejemplos: PostGIS (software libre), Oracle Locator y Spatial (propietario)

Programación en Python o R

- Estos dos grandes sistemas para ciencia de datos también pueden operar con IG gracias a diversas librerías
- Aporta más escalabilidad horizontal que las BDRG, ya que los algoritmos complejos pueden fragmentarse y paralelizarse
- Todos los algoritmos de la GIScience pueden reproducirse sin cambios (escalabilidad vertical) o modificarse para paralelizarlos (escalabilidad horizontal)
- Alta expresividad para crear nuevos algoritmos
- Proporcionan todo el extenso catálogo de librerías de estos sistemas para combinarlos con el análisis específico de IG, por ejemplo, su rica funcionalidad de Machine Learning
- Ejemplos (todo software libre): GDAL / OGR, Fiona, Shapely, Rasterio, GEOS

Programación en stacks distribuidos

- El stack más utilizado por ahora es Spark, puesto que puede usarse con Python (PySpark) y por tanto aprovechar las librerías anteriormente comentadas

- Todos los algoritmos clásicos de GIScience tienen que ser modificados para adaptarse a este entorno de computación distribuida

- Las soluciones genéricas son complejas debido a la alta interrelación de la IG debido a la topología, se tiende a crear un programa específico para cada problema

SaaS

- Existen productos SaaS orientados al análisis de datos espaciales
- Orientados al usuario final:
  - Casi todas las plataformas de BI ofrecen algún nivel de funcionalidad espacial: Tableau, Microsoft Power BI, IBM Watson, etc., pero no incorporan algoritmos GIScience complejos
  - CARTO: es una plataforma que ofrece, sobre distintos almacenes de datos como Google BigQuery, analíticas y visualización de información geográfica muy avanzadas
  - ArcGIS Online: la versión cloud de ArcGIS lleva la experiencia del GIS de escritorio clásico a la nube
  - Google Earth Engine: orientado a teledetección y Earth Observation Sciences, permite acceder a vastos repositorios de imágenes multiespectrales de satélite y programar en Python algoritmos de procesamiento

SaaS

- Orientados a desarrolladores:
  - Google: Google Maps y sus productos satélites: cartografía, navegación en rutas, acceso a datos geolocalizados recogidos por Google, Google Engine para llevar a cabo masivos flujos de trabajo con información raster;
  - Mapbox: múltiples APIs de servicio para acceder a cartografía personalizada, navegación en ruta, datos espaciales reales para videojuegos, análisis de geomárketing, visualización 3D, etc.

03

Caso 00
Meteorología para la navegación aérea

Caso de uso: Meteorología para navegación aérea

El reto: hacer el Big Data lo suficientemente Small para que quepa en un iPad

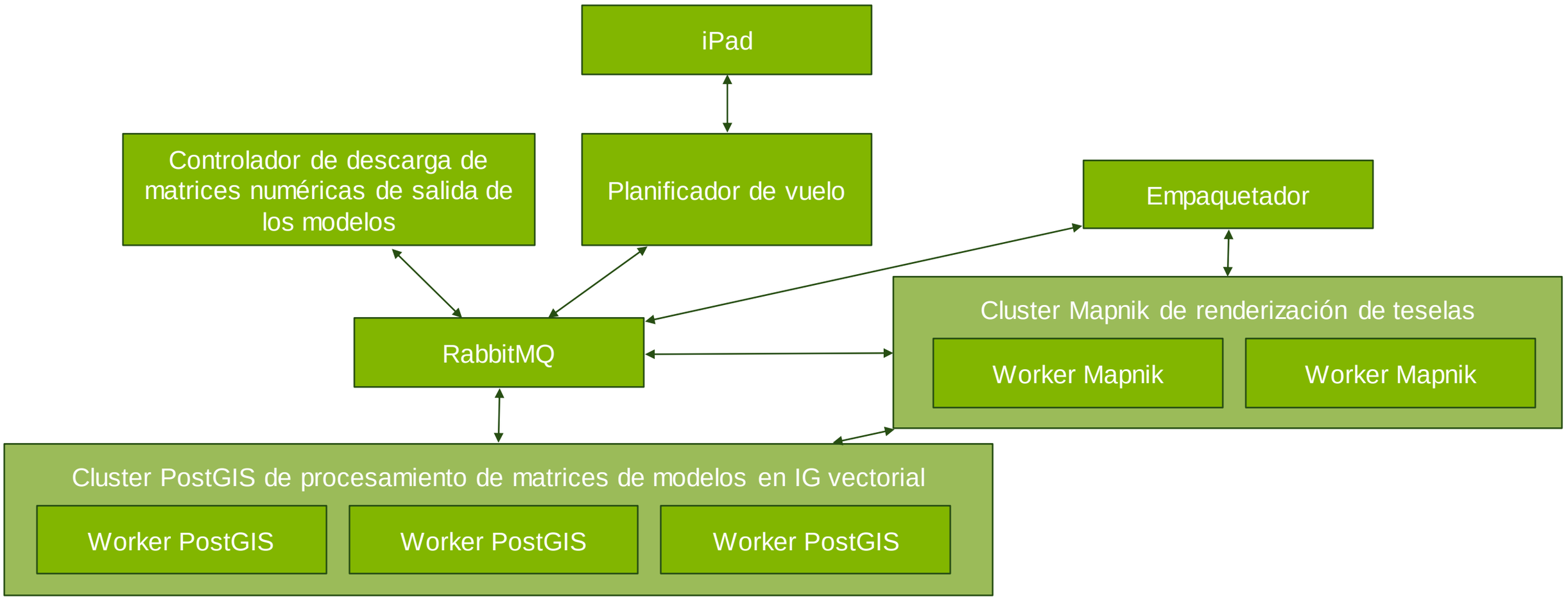
- La planificación de los vuelos comerciales utiliza una gran cantidad de datos meteorológicos proporcionados por agencias gubernamentales especializadas
- Entre las variables necesarias para el vuelo está desde la topografía de la tropopausa hasta los avisos de difusión de cenizas volcánicas, pasando por las zonas de turbulencias, temperatura a cada nivel de vuelo, etc.
- Todos estos datos hay que multiplicarlos por los diversos momentos temporales que los modelos meteorológicos generan, en función de la duración del vuelo

Caso de uso: Meteorología para navegación aérea

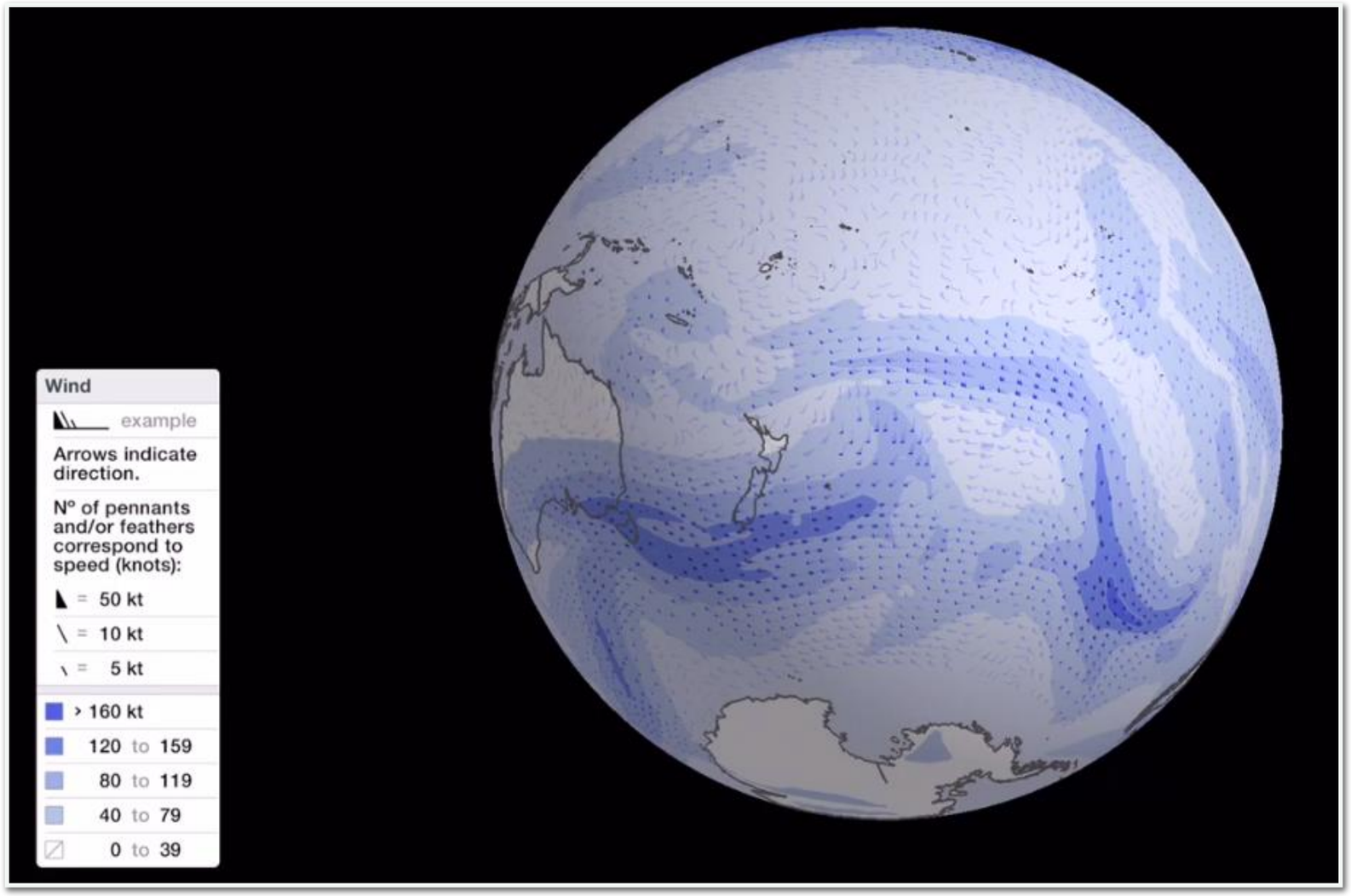
El reto: hacer el Big Data lo suficientemente Small para que quepa en un iPad

- Tradicionalmente, toda esta información se lleva a la cabina en forma de papel
- Actualmente, el uso de Internet en las cabinas no está generalizado
- Por lo tanto, si se quiere acceder a toda esta información en un dispositivo portátil, hay que hacerlo en condiciones de estricta desconexión, disponiendo sólo de los recursos del dispositivo en cuestión, en este caso, un iPad sin modificación alguna
- La solución: crear una plataforma de computación en la nube que prepare y empaquete los datos para los pilotos de forma que éstos puedan descargarla a sus dispositivos antes de quedar desconectados

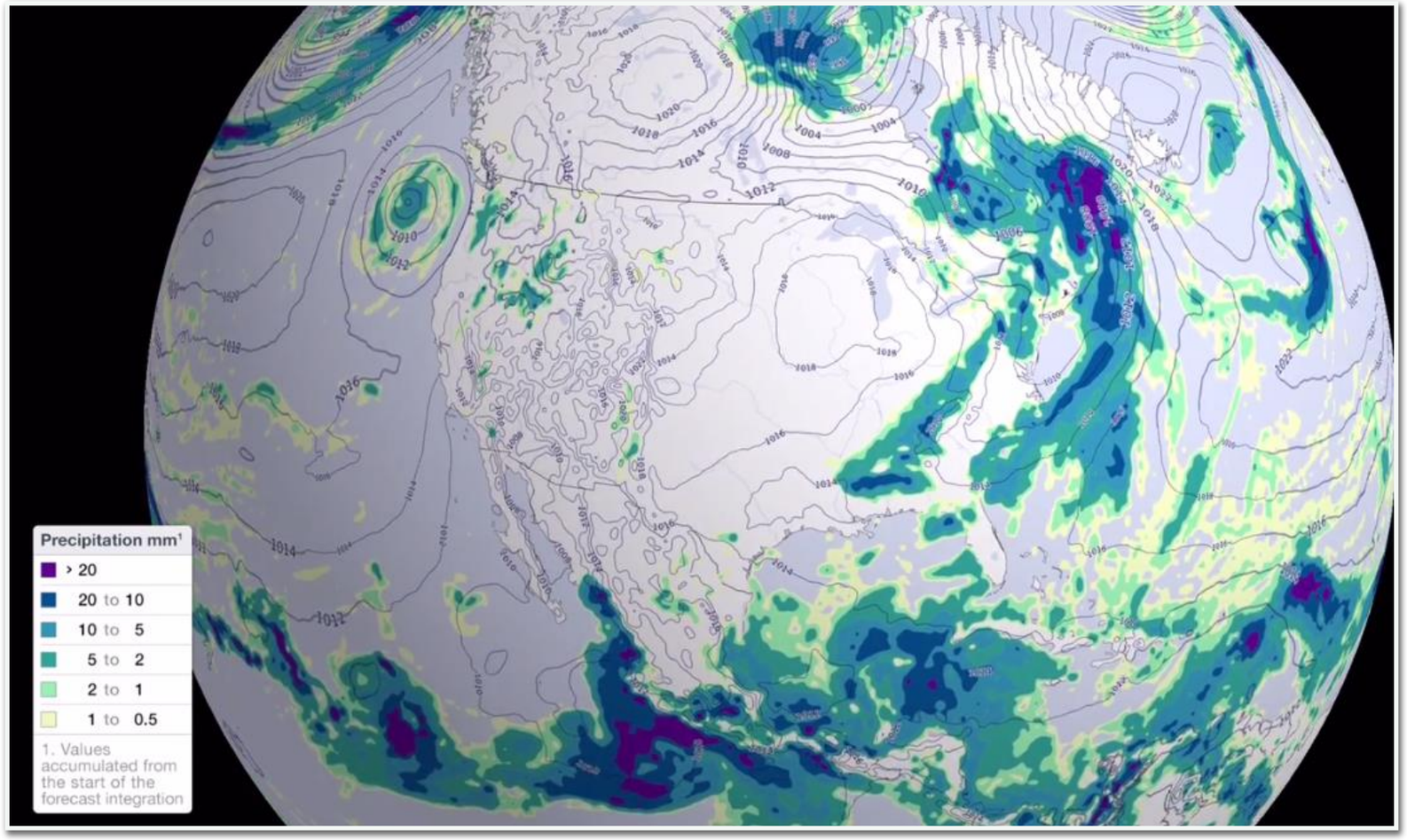
Caso de uso: Meteorología para navegación aérea



Caso de uso: Meteorología para navegación aérea

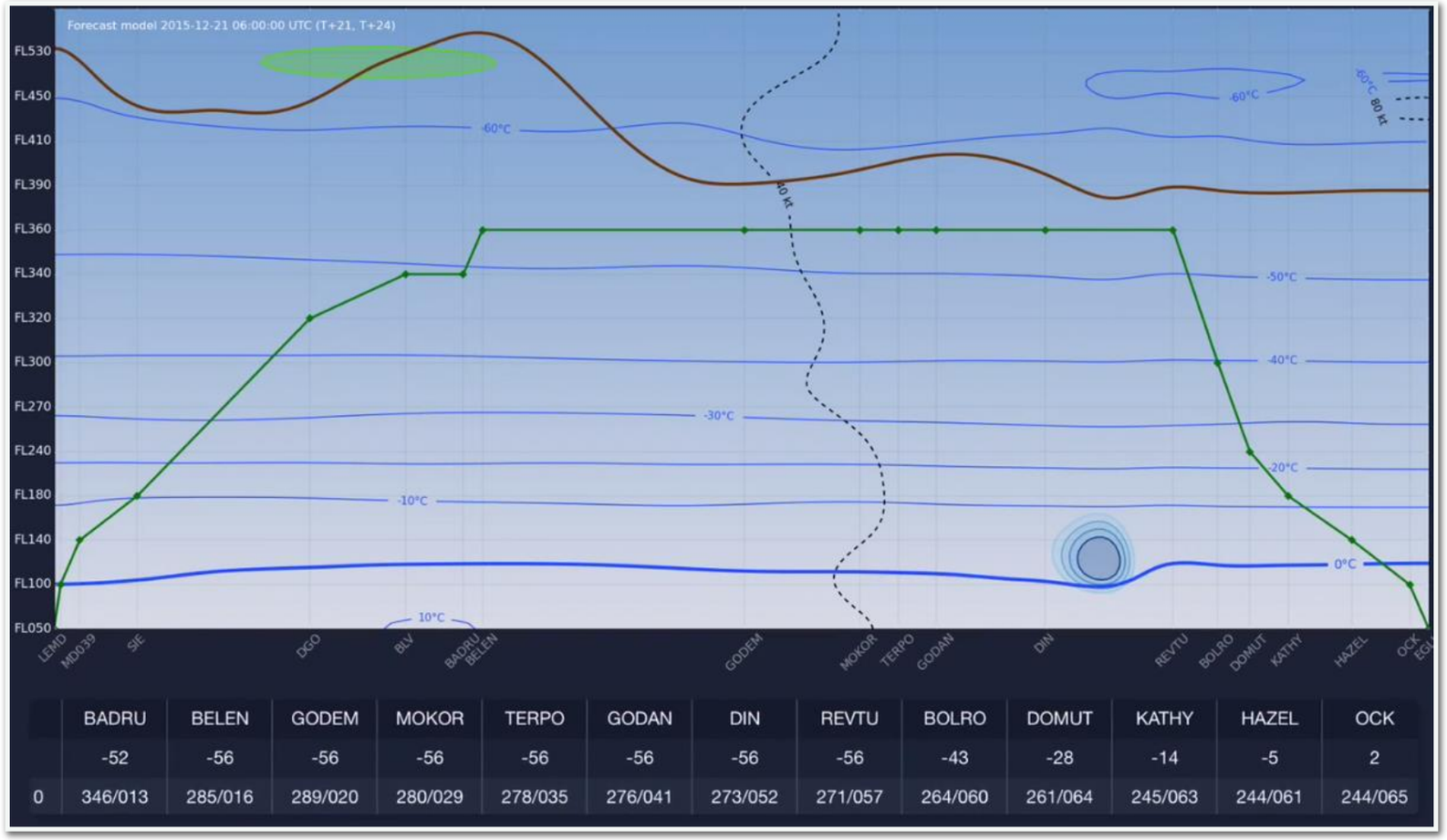


Caso de uso: Meteorología para navegación aérea



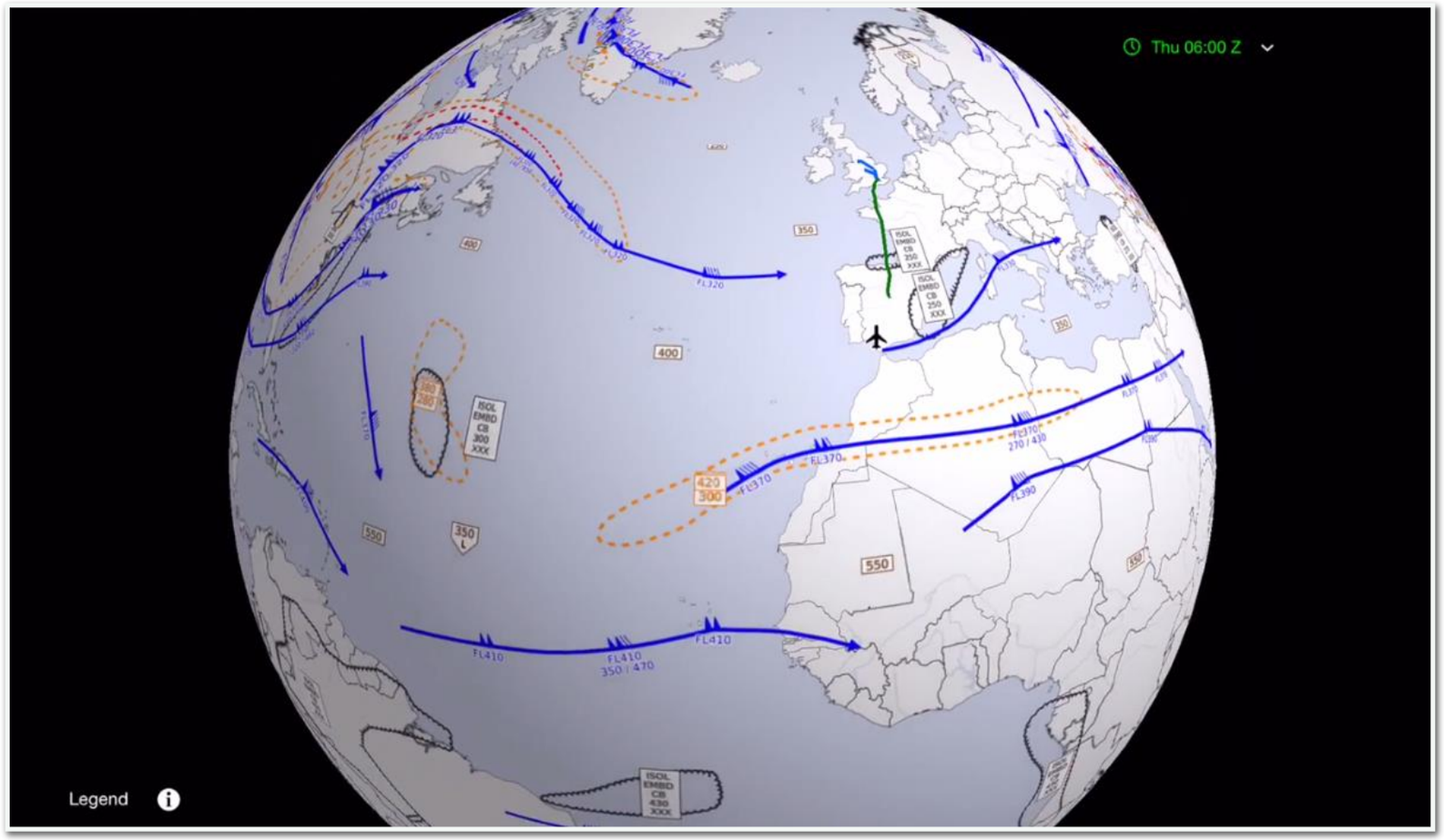
Caso de uso: Meteorología para navegación aérea

Caso de uso: Meteorología para navegación aérea



	BADRU	BELEN	GODEM	MOKOR	TERPO	GODAN	DIN	REVTU	BOLRO	DOMUT	KATHY	HAZEL	OCK
	-52	-56	-56	-56	-56	-56	-56	-56	-43	-28	-14	-5	2
0	346/013	285/016	289/020	280/029	278/035	276/041	273/052	271/057	264/060	261/064	245/063	244/061	244/065

Caso de uso: Meteorología para navegación aérea



04

Caso 01
Climatología. Análisis de tendencias y escenarios de cambio climático

Caso de uso: Climatología. Análisis de tendencias y escenarios de cambio climático

Existe una enorme cantidad de información científica de primera calidad de acceso público, el problema es que hacerla accesible y comprensible por los no especialistas no siempre es una prioridad

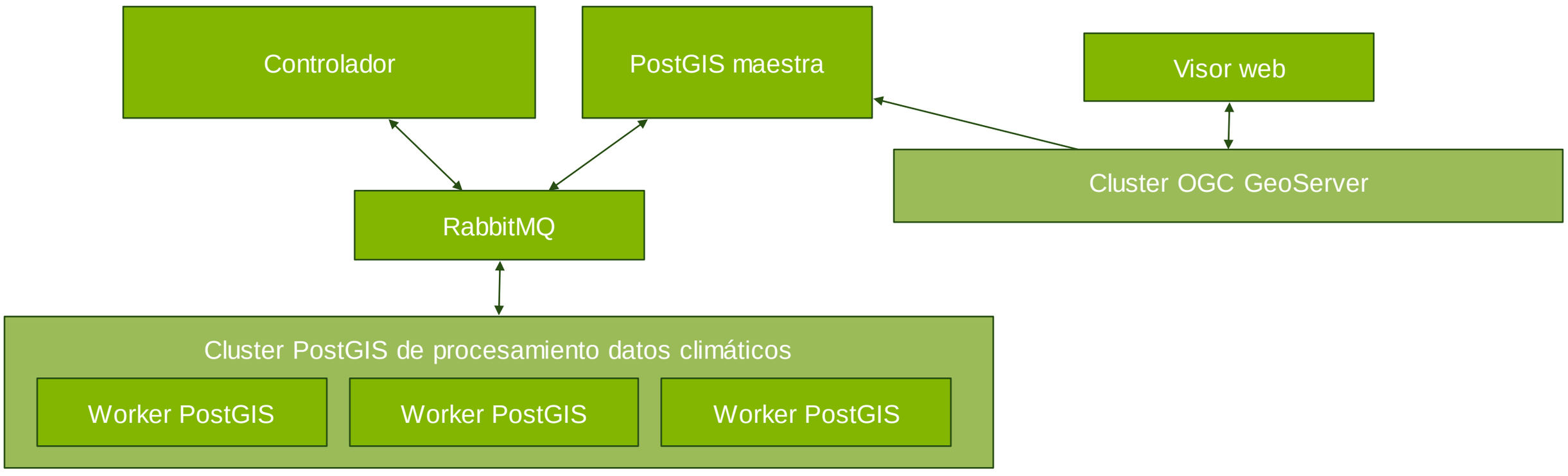
- Y eso es precisamente lo que pretende globalclimatemonitor.org: hacer entendible por el público en general la información de anomalías climáticas que procesa y publica la Climate Research Unit (CRU) de la Universidad de East Anglia, uno de los centros de referencia mundiales en cuestiones de cambio climático
- La CRU pone a disposición pública su enorme banco de datos climáticos históricos
- Es un proyecto del Grupo de Investigación del Clima del Dpto. de Geografía Física de la Universidad de Sevilla

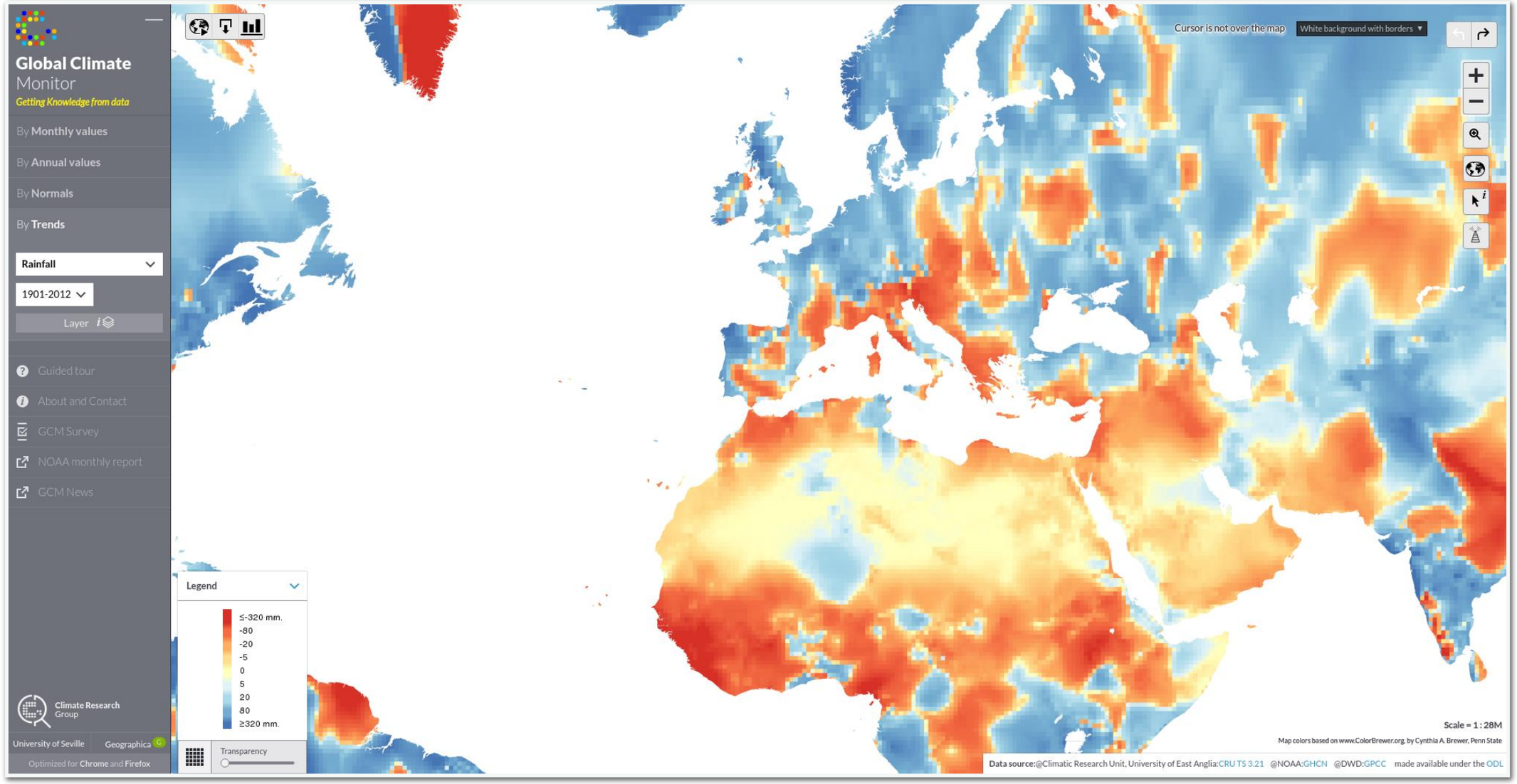
Caso de uso: Climatología. Análisis de tendencias y escenarios de cambio climático

Una vez más, se trata de plasmar en datos procesados usables e interpretables un conjunto de datos de tamaño moderado y procesamiento geográfico intensivo

- Actualmente, globalclimatemonitor.org utiliza aproximadamente 50 GB de datos climáticos almacenados en una instancia PostGIS escalada verticalmente y organizados en tablas que alcanzan los 90 millones de registros para algunas variables
- A estos datos se les realizan diversos tratamientos de análisis climático para estimar la proyección de cambio de diversas variables y las anomalías registradas en el histórico

Caso de uso: Climatología. Análisis de tendencias y escenarios de cambio climático




Global Climate
Monitor
Getting Knowledge from data
By Monthly values
By Annual values
By Normals
By Trends
Rainfall
1901-2012
Layer
Guided tour
About and Contact
GCM Survey
NOAA monthly report
GCM News
Climate Research Group
University of Seville
Geographica
Optimized for Chrome and Firefox
Legend
≤-320 mm.
-80
-20
-5
0
5
20
80
≥320 mm.
Transparency
Cursor is not over the map
White background with borders
Scale = 1 : 28M
Map colors based on www.ColorBrewer.org, by Cynthia A. Brewer, Penn State
Data source:@Climatic Research Unit, University of East Anglia:CRU TS 3.21 @NOAA:GHCN @DWD:GPCC made available under the ODL