esri Canada

# Big data, big deal?

Presented by:
David Stanley, CTO
PCI Geomatics

# A quick word about me…

- Ah, the early 1980s

# Aha! Moments in my Career

- First personal computer (Pet 2001, 1978)
- First windowing OS (SUN, ~1985)
- First Web Browser (Mosaic 1992)
- Mobile computing – (iPhone 2007)



And now…The Cloud

# What's the Cloud?

# What about Esri Maps?

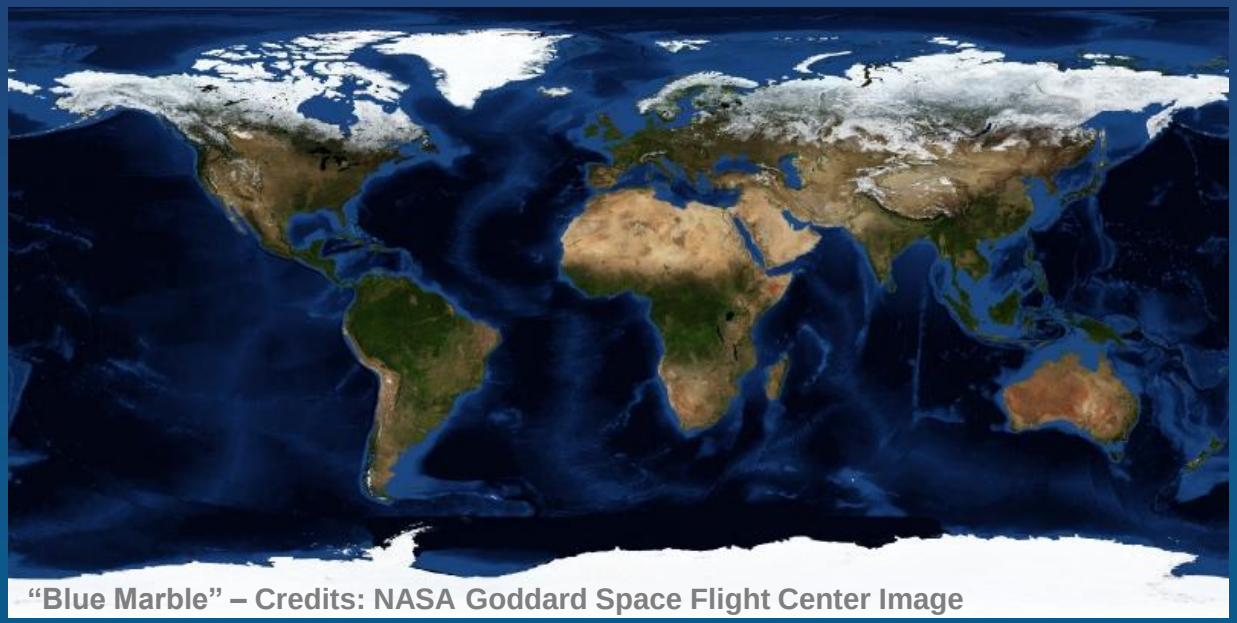
- Esri recognized the need for high quality imagery on its online platform “ArcGIS Online”
- Imagery is a key component of Esri’s new Cloud Services – providing high resolution, high quality coverage is very important
- Wanted a comprehensive 1m imagery base map
- Not an easy thing to do – witness Apples recent iOS 6 map application problems

# The Project: King Kong

- In 2011, Esri approached PCI to assist with given its 30 year history in image processing and high capacity computing
- Project – named "King Kong", represented a considerable challenge for both Esri and PCI
  - Huge data volumes
  - Large data management issues
  - Massive processing requirements
  - Difficult, finicky algorithms
- One Year (12 months) time frame

# ‘Big Data’

- 50 million $km^2$ of imagery at 1m resolution covering world
- 300,000+ IKONOS Scenes
- Hundreds of Terabytes of raw input data
- Thousands of Terabytes (Petabytes) in intermediate data



“Blue Marble” – Credits: NASA Goddard Space Flight Center Image

# The Traditional Approach

- Find a large amount of space
- Purchase 200+ compute servers
- Purchase farm of network data servers
- Setup a complex network/infrastructure
- Hire 30+ person team for processing, QA and IT

- After project done in 12 months try to figure out what to do with it all…

# The Bold Approach

- Small permanent team : 4 from  ESRI, 1 from PCI
- Rely on enormous blocks of imagery (1000's) and huge number of tie-points (100,000's) to average out errors in colour and geometry to dramatically reduce labour.
- Rely on massive parallelism for computation and speed
- Gamble on clever people and algorithms
- Go with the Amazon Cloud!

# Why consider the cloud?

| Reason | Description |
|---|---|
| High Availability | Always on |
| Highly Accessible | Access via any web portal/browser |
| Scalable | Expand or contract processing needs as required. Can get 100x more power for a while, then let go. |
| Fault Tolerant | 1000's of shared servers, a few going down is not even noticed. |
| Running costs | Pay as you go for what you actually use.<br>No processing = no cost. |
| Avoid Capital outlays | No need to buy hardware and maintain it. |
| The Bad thing though…<br>Data? | Moving large amounts of imagery on and off the cloud can be challenging, slow and perhaps costly… |

# Amazon Cloud Components

Simple Storage Service (S3) Buckets

Elastic Block Storage (EBS) Volumes (1 TB Max)

EBS Snapshot (S3)


Amazon Machine Instance (AMI)

Config Operation

Source Data Transport

Scaled Instances

Internet

S3 Object (5GB Limit)


Object Key

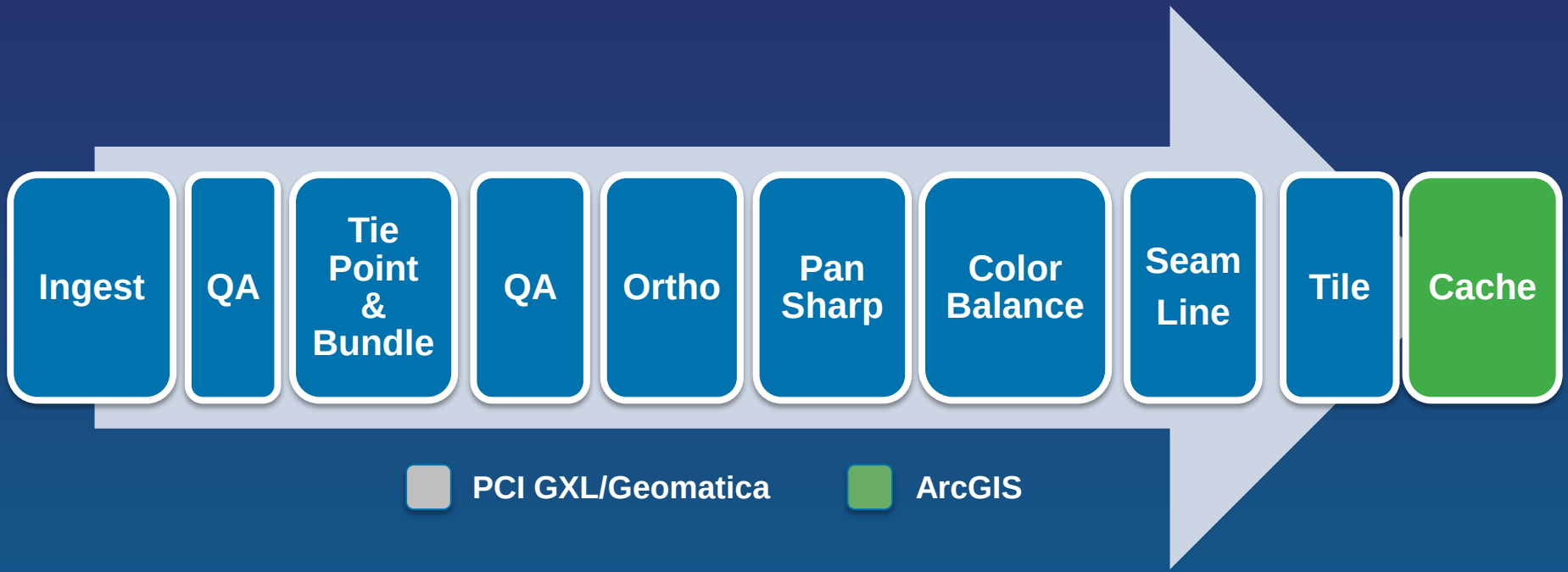
# “Standard” compute node

**“Standard” Compute Node**

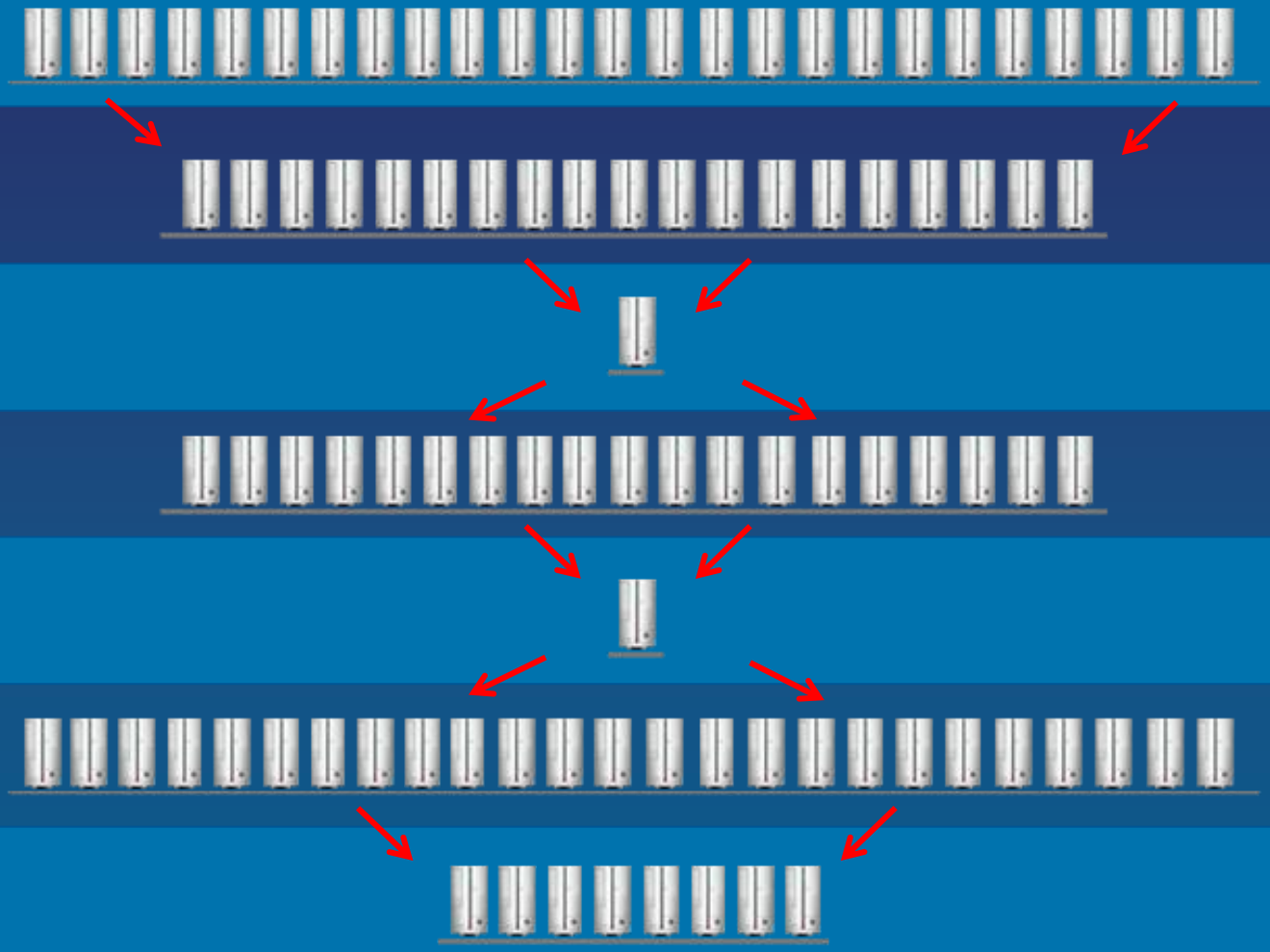
- 15 GB memory
- 8 EC2 Compute Units (4 virtual cores with 2 EC2 Compute Units each)
- 1,690 GB ephemeral storage (“free”)
- 8 GB EC2 Volume
- 64-bit Linux platform
- I/O Performance: High

**Out of many available configurations this was best**

# Processing workflow



# Why the cloud worked



Ingest
(25+ Nodes)

Tie Point
(20+ Nodes)

Bundle
(1 Node)

Ortho / Pan Sharpen
(20+ Nodes)

Color Balance
(1 Node)

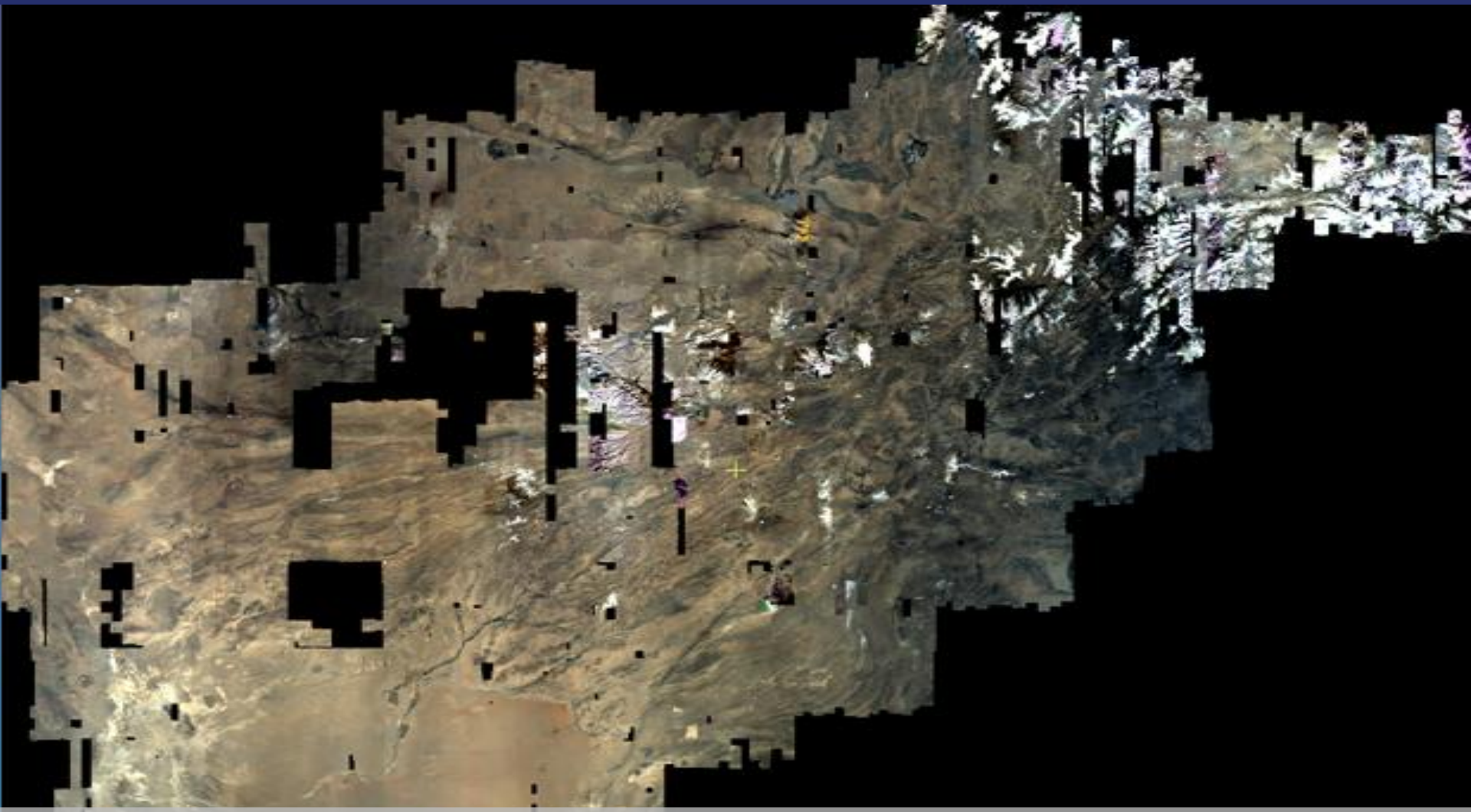
Tiling
(25+ Nodes)

Caching
(8+ Nodes)

# The Project in Action

- 1st 6 months: getting it working
  - Struggled with our own software when scaling from hundreds to thousands
  - Struggled with finding the ‘right’ cloud configuration for computing and storage
  - Attempting to reduce, or automate, manual QA steps
- 2nd 6 months: crank it up
  - At the end processing 70,000+ scenes per month

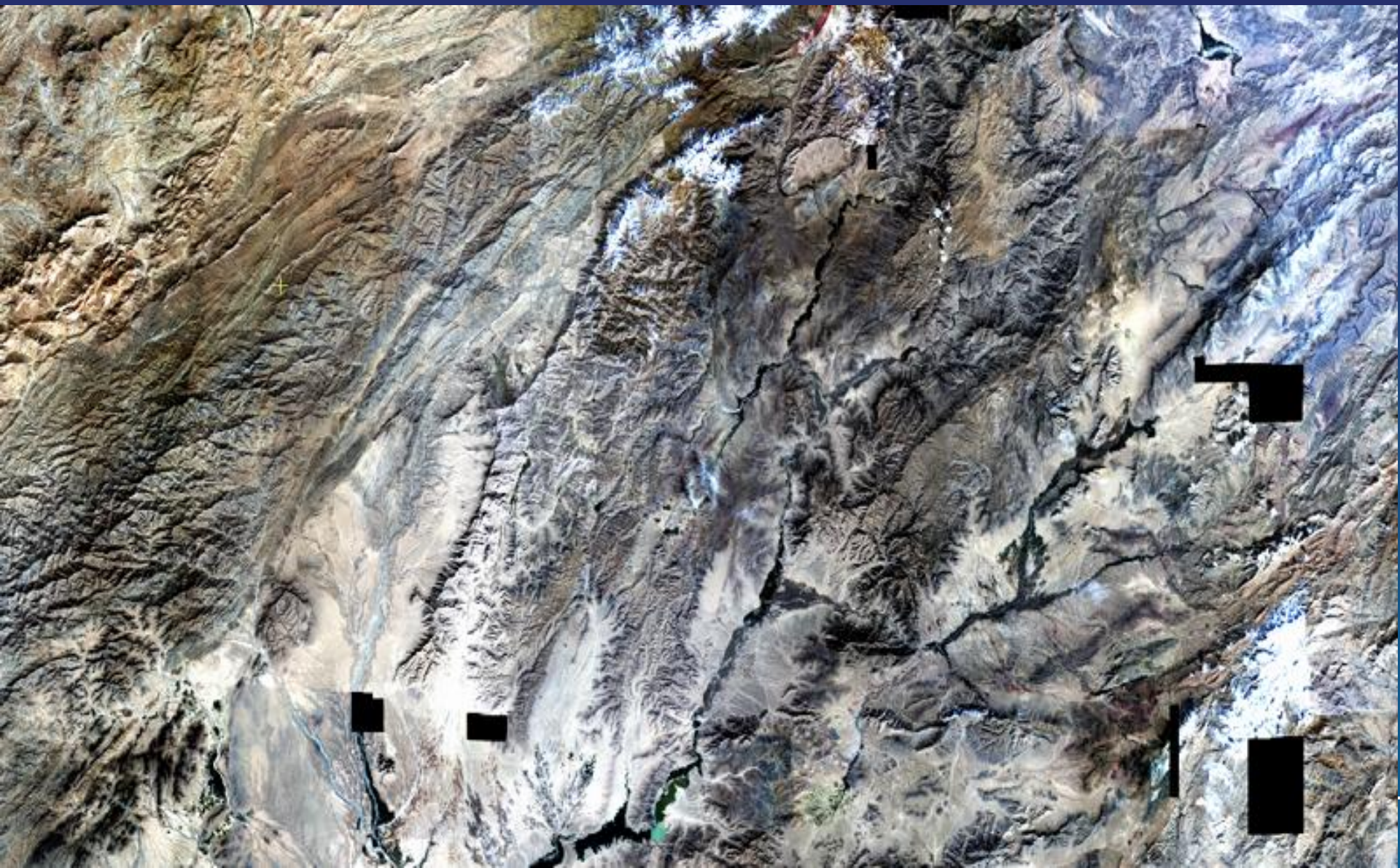
# A big data set: Afghanistan



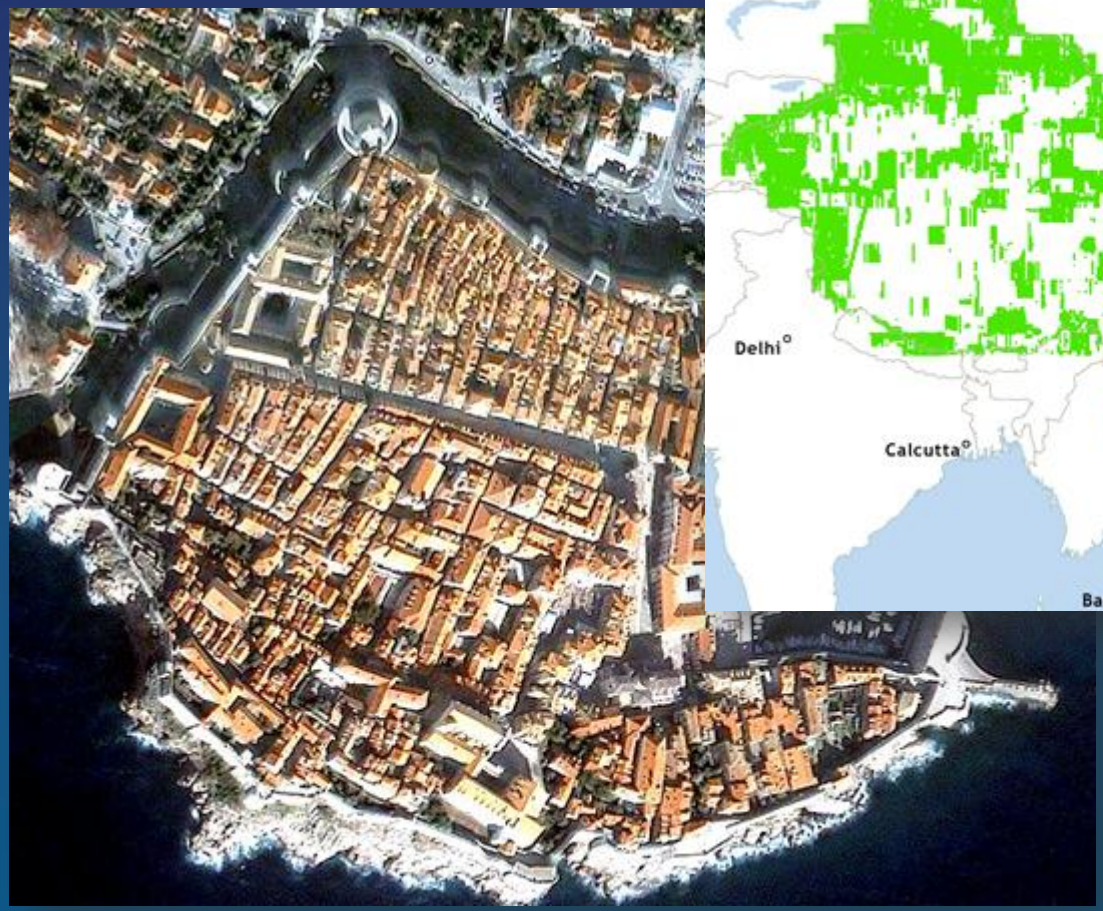
9,500 Ikonos scenes. 15TB of Data. 1,000,000+ $km^2$

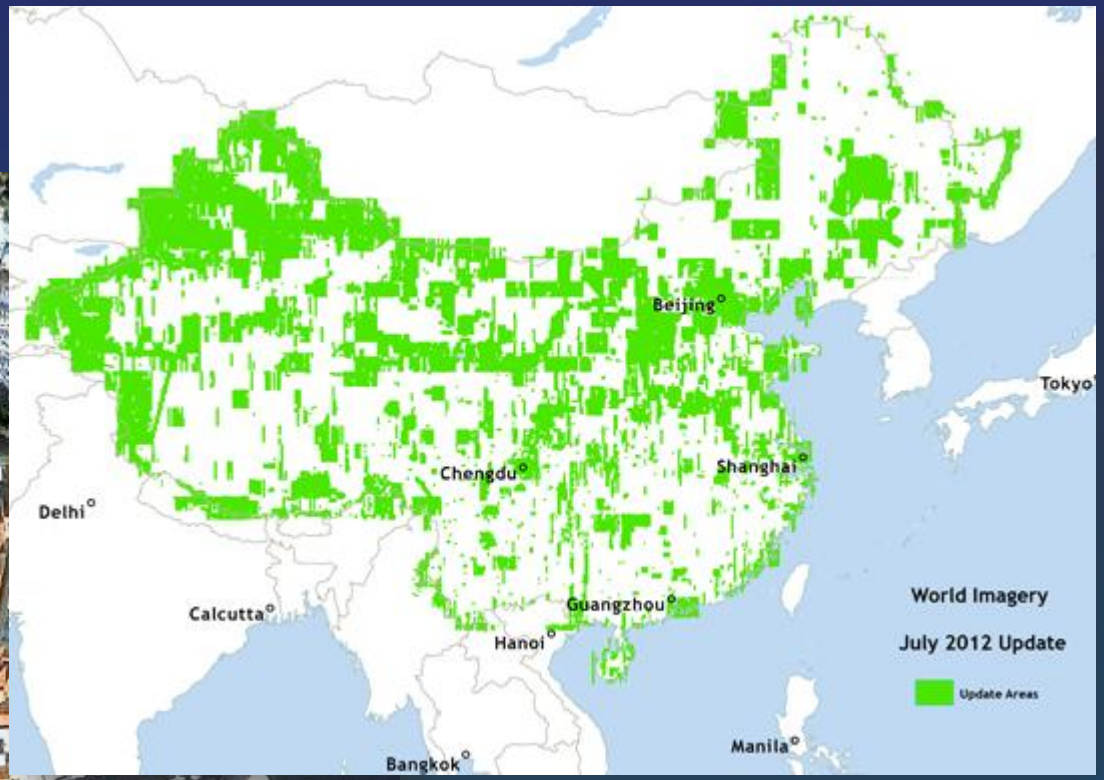
# Colour Balancing: Before

# Colour Balancing: after

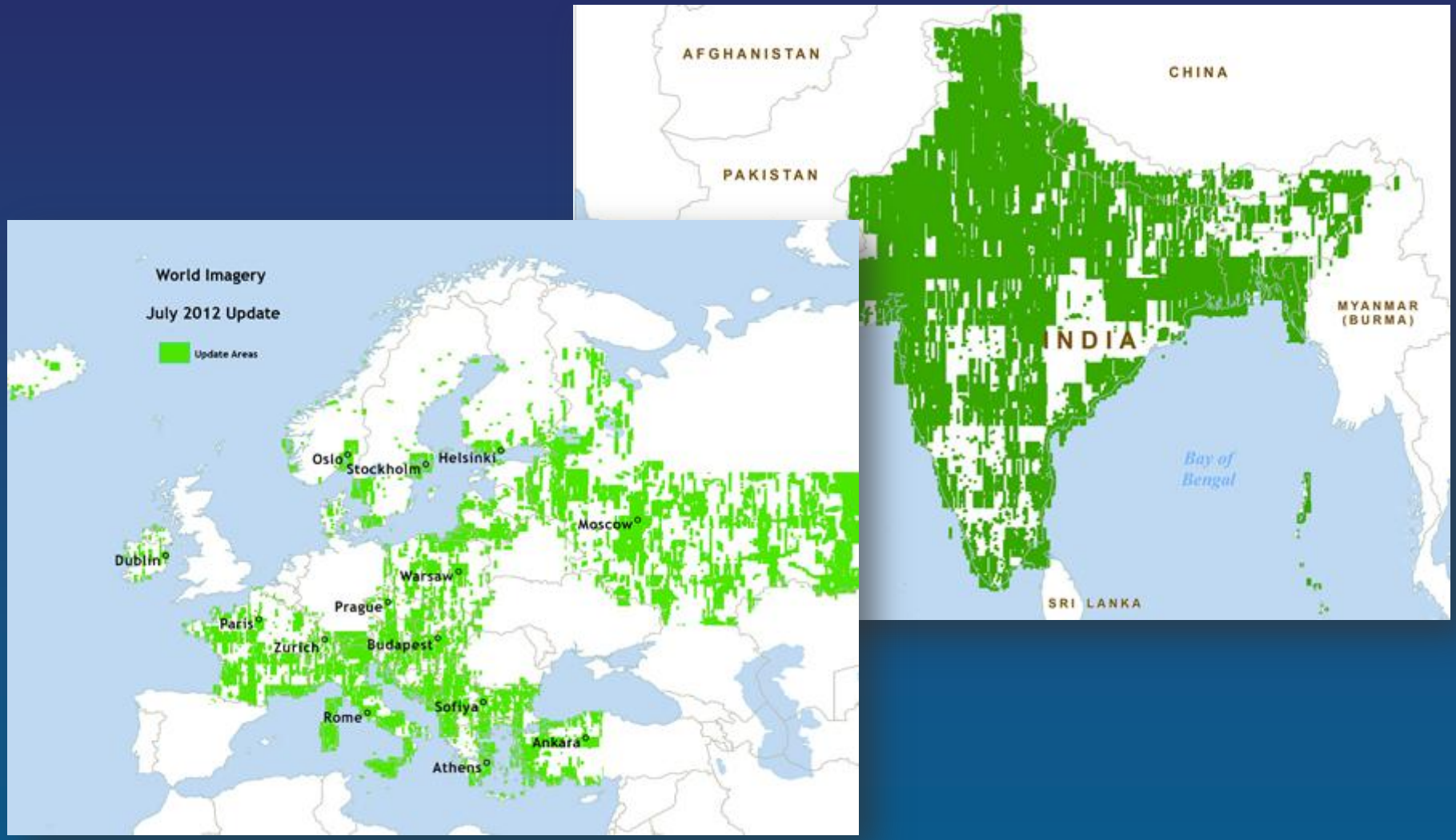


# World Coverage – 50 million km2





# World Coverage – Examples

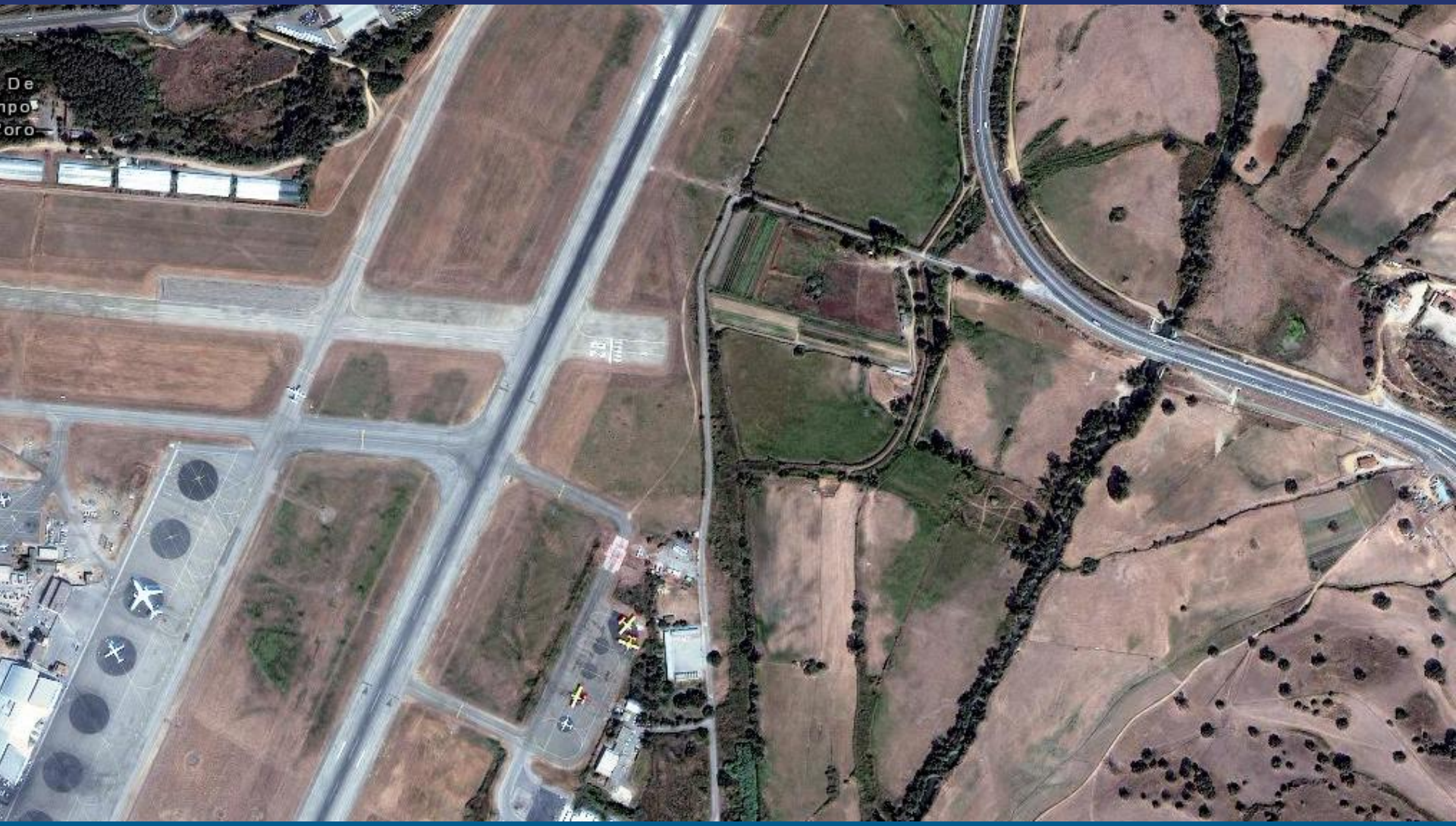


# Sample output

- Taj Mahal, India

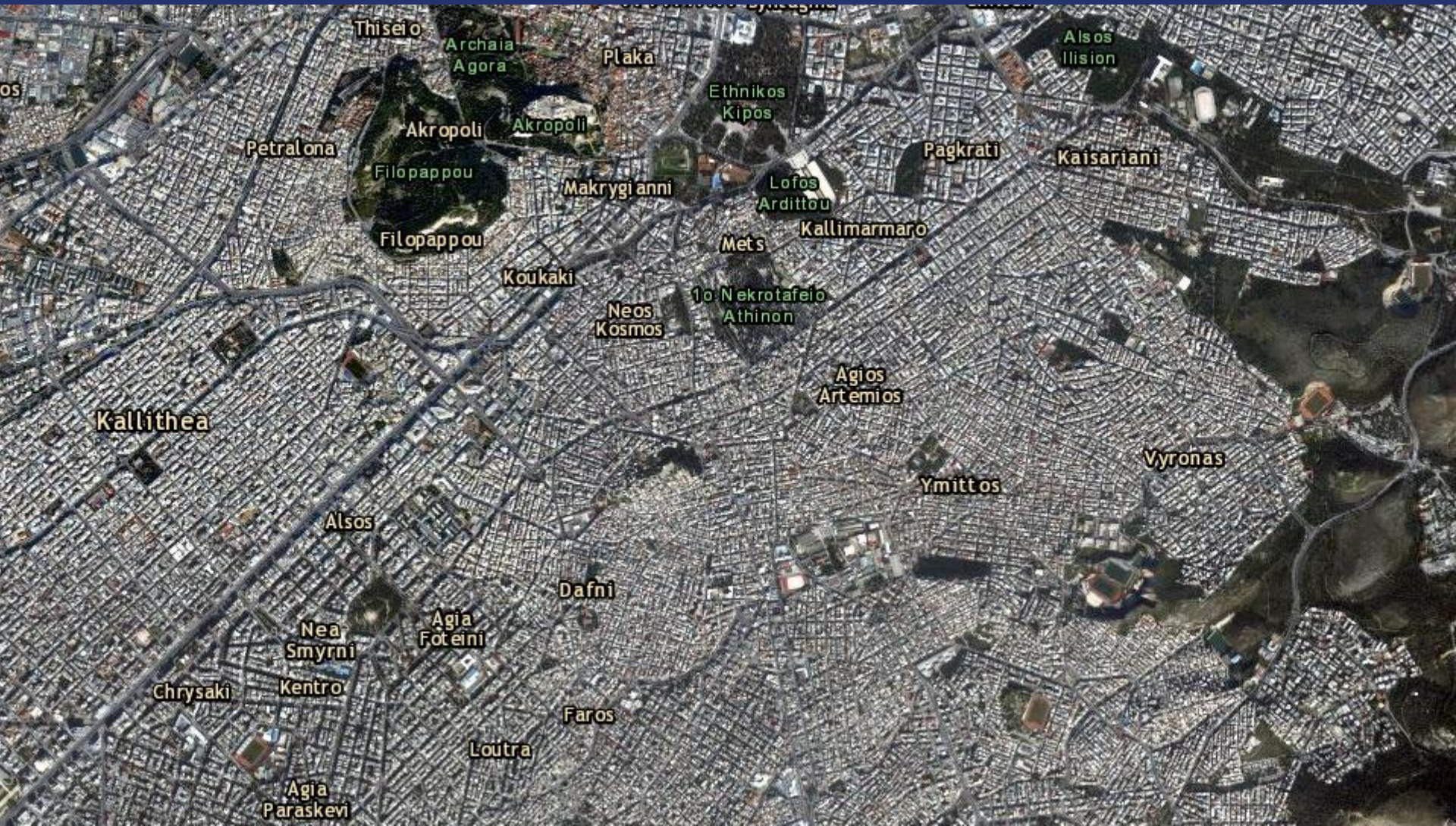
# Sample output

- Corsica, France



# Sample output

- Athens, Greece



# Sample output

- Abu Dhabi, UAE

# Summary

- Approximately 4m accuracy on the ground
- Alignment between images good
- Now that done, could probably redo in 3 months.
- The cloud is effective!

As with most leading edge projects: it was an amazing, stressful and fun experience.

Enjoy the new 1m image data layers!

# Q & A


## THANK YOU

**David Stanley, CTO**
**PCI Geomatics**
**stanley@pcigeomatics.com**