GIS for Big Data

Lau Wee Lik

What is GIS

- A geographic information system (GIS) lets us visualize, question, analyze, and interpret data to understand relationships, patterns, and trends.
- GIS benefits organizations of all sizes and in almost every industry. There is a growing interest in and awareness of the economic and strategic value of GIS.

Age of Data Ubiquity

- Data is now central to our existence – both for corporations and individuals

- Nimble, thin, data-centric apps exploiting massive data sets generated by both enterprises and consumers

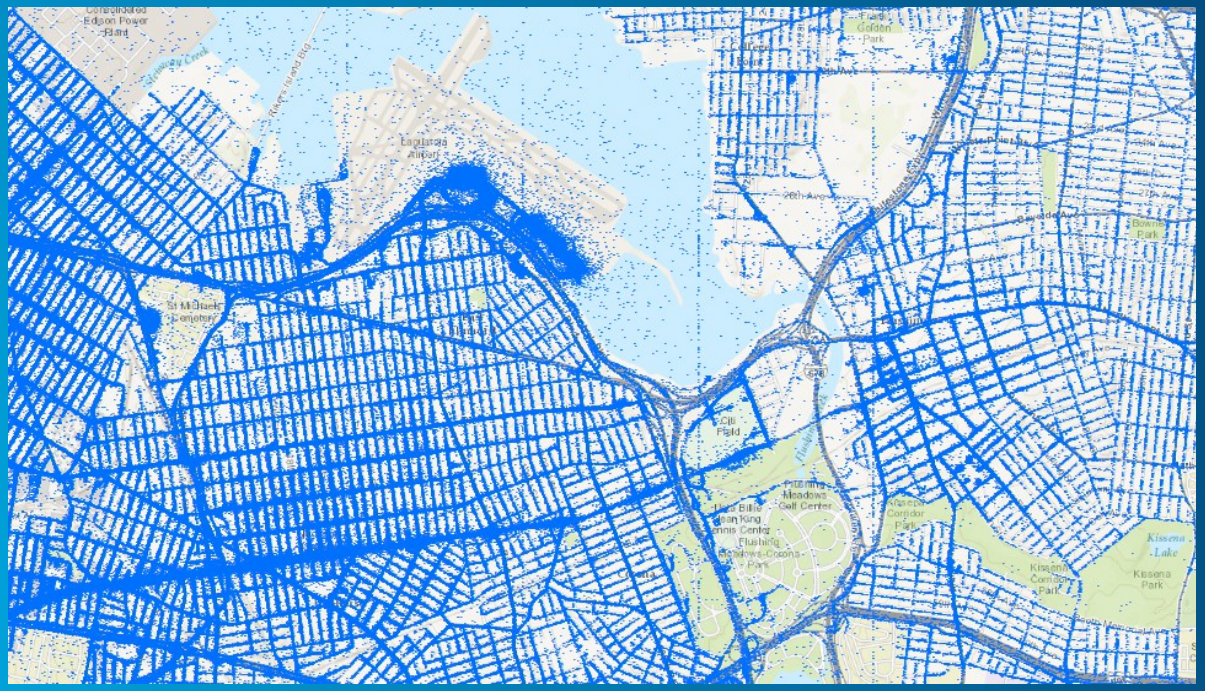
- Hardware era: 20 – 30 years
- Software era: 20 – 30 years
- Data era: ?

Big data
What is it?

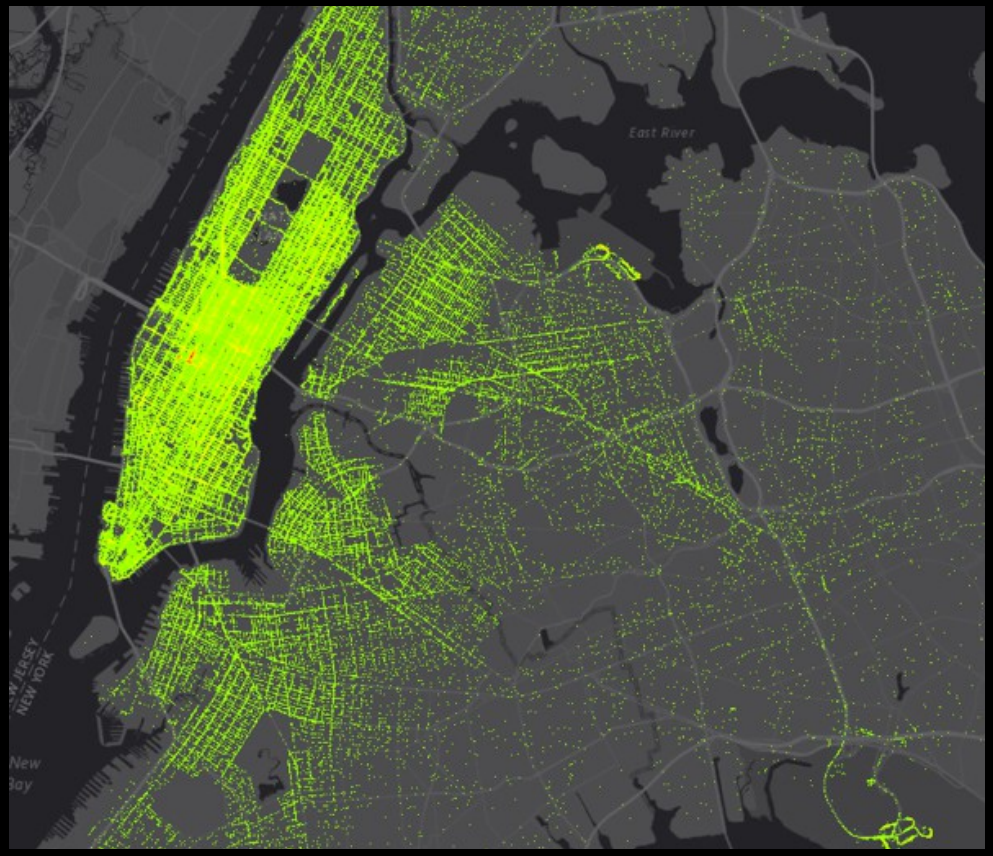
- **Big Data is a loosely defined term used to describe data sets so large and complex that they become awkward to work with using standard software in a tolerable elapsed time**
  - **Big data "size" is a constantly moving target, ranging from a few dozen terabytes to many petabytes of data**
  - **In the past three years, 90% of all recorded data has been generated**
- **Every 60 seconds:**
  - **100,000 tweets**
  - **2.4 million Google searches**
  - **11 million instant messages**
  - **170 million email messages**
  - **1,800 TB of data**

NYC Taxi data (2013) – 170 millions trips

- **Users have huge quantities of valuable data but are having a hard time dealing with it**
  - **Hard to manage**
  - **Struggle to visualize**
  - **Unsure of what is interesting what questions to ask and what to analyze.**



NYC Taxis by Day



Manhattan Taxis Friday after 8pm

GIS users have big data

- Smart Sensors
  - Electrical meters (AMI), SCADA, UAVs
- GPS Telemetry
  - Vehicle tracking, smartphone data collectors, workforce tracking, geofencing
- Internet data
  - Social media streams, web log files, customer sentiment
- Sensor data
  - Weather sensors, stream gauge measurements, heavy equipment monitors, …
- Imagery
  - Satellites, frame cameras, drones

Value when analyzing data at mass scale

- As observations increase in frequency
  - Each individual observation is worth less
  - …as the set of all observations becomes more valuable

- One single metric from the jet aircraft is much less useful than the analysis of that metric against the same metric from every known flight of that aircraft over time

- *Big Data* is the accumulation and analytical processes that uses this data for business value

Big challenges

- Data acquisition
  - Filtering and compressing
  - one million terabytes per day
- Information extraction and cleaning
- Data integration, aggregation, and representation
  - Heterogeneous datasets
- Modeling and analysis
  - Nonstandard statistical analysis; very noisy, dynamic, and untrustworthy
- Interpretation
  - Decision making – metadata, assumptions, very complex

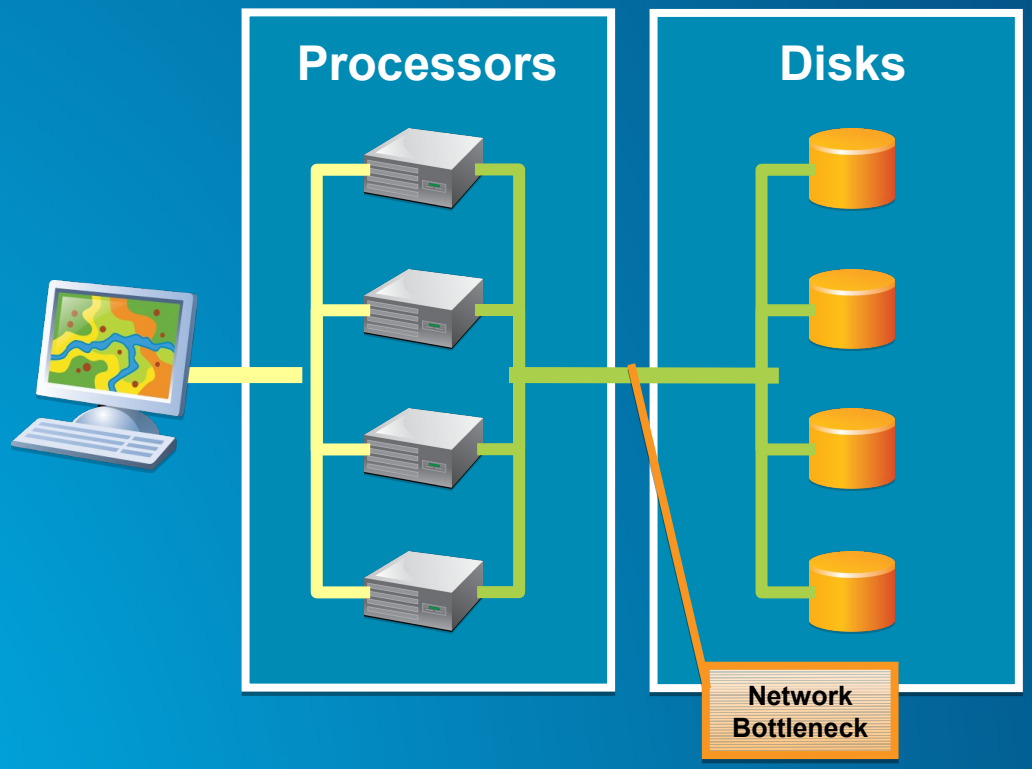
Big data

What techniques are applied to handle it?

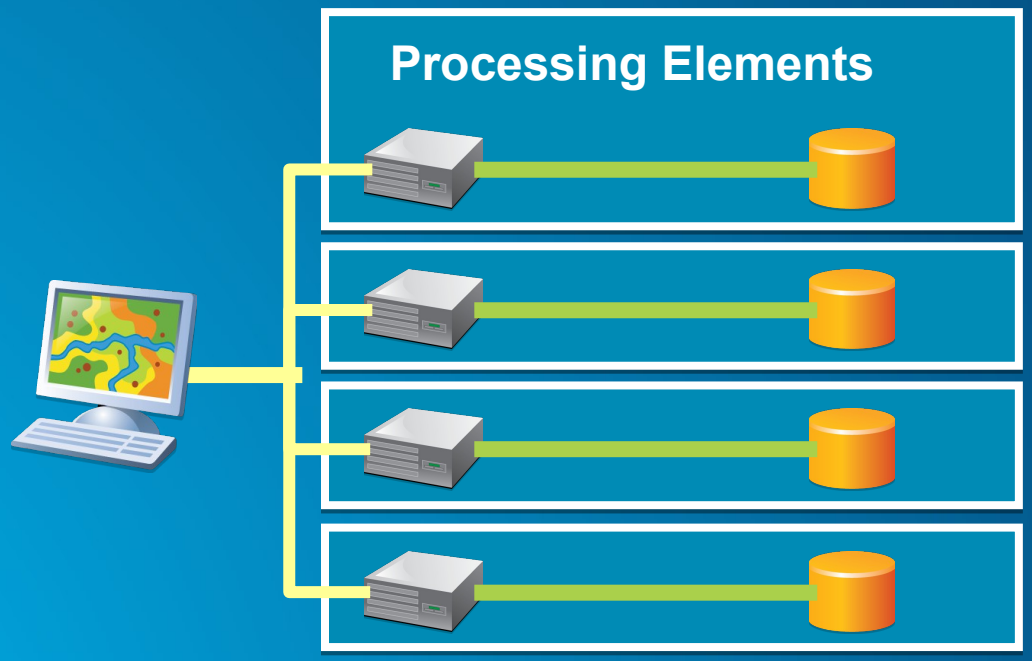
- **Data distribution** – large datasets are split into smaller datasets and distributed across a collection of machines
- **Parallel processing** – using a collection of machines to process the smaller datasets, combining the partial results together
- **Fault tolerance** – making copies of the partitioned data to ensure that if a machine fails, the dataset can still be processed
- **Commodity hardware** – using standard hardware that is not dependent upon exotic architectures, topologies, or data storage (e.g., RAID)
- **Scalability** – algorithms and frameworks that can be easily scaled to run on larger collections of machines in order to address larger datasets

“Hadoop” Distributed File System

Legacy system architecture

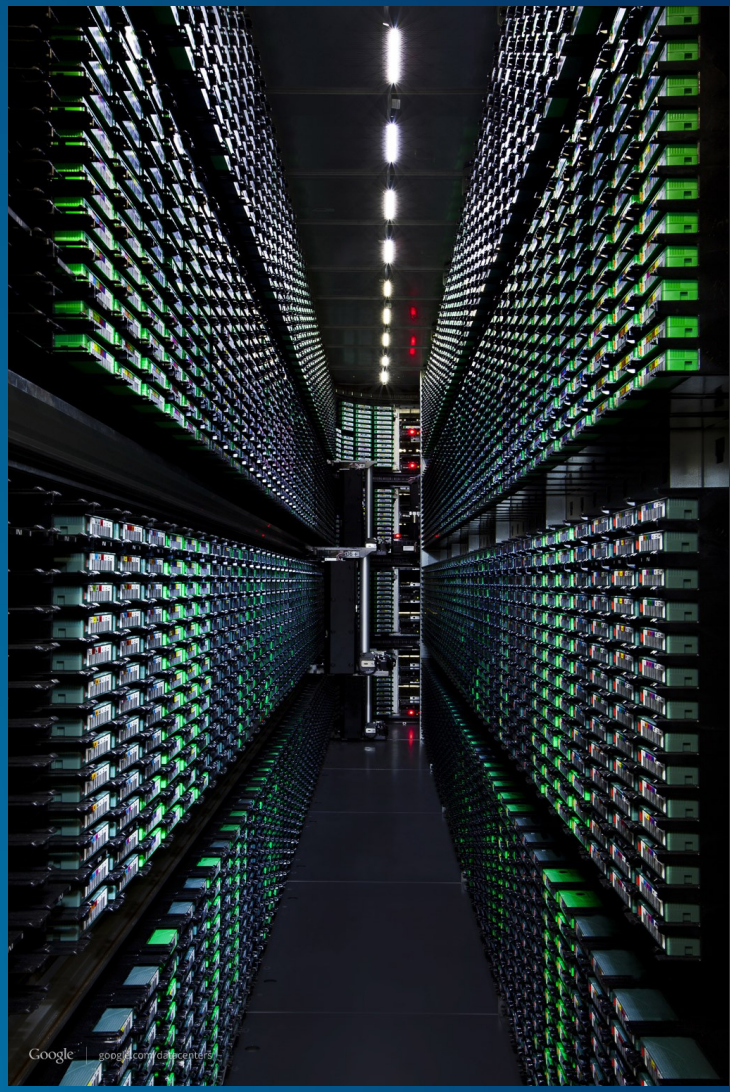


Distributed system architecture



Until Now…

- Google implemented their enterprise on a distributed network of many nodes, fusing storage and processing into each node
- Hadoop is an open source implementation of the framework that Google has built their business around for many years



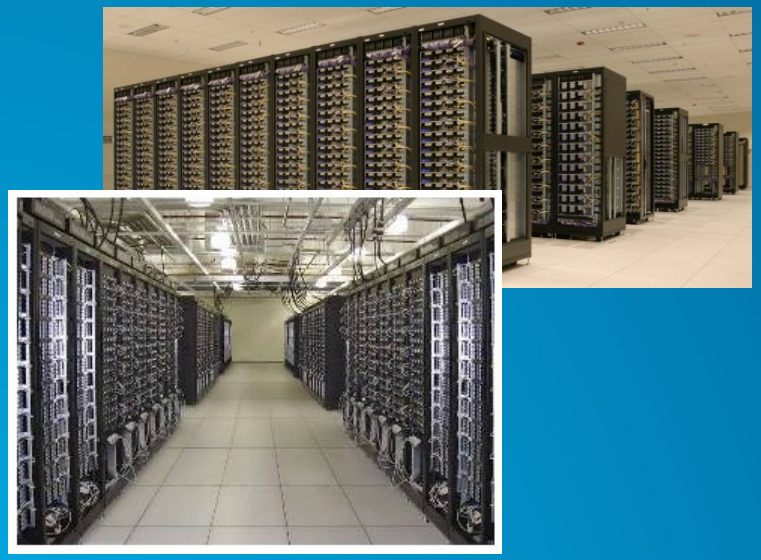
Apache Hadoop

Overview

- Hadoop is a scalable open source framework for the distributed processing of extremely large data sets on clusters of commodity hardware
  - Maintained by the Apache Software Foundation
  - Assumes that hardware failures are common
- Hadoop is primarily used for:
  - Distributed storage
  - Distributed computation

Apache Hadoop

Hadoop Clusters



Traditional Hadoop Clusters

20 flowdown PCs

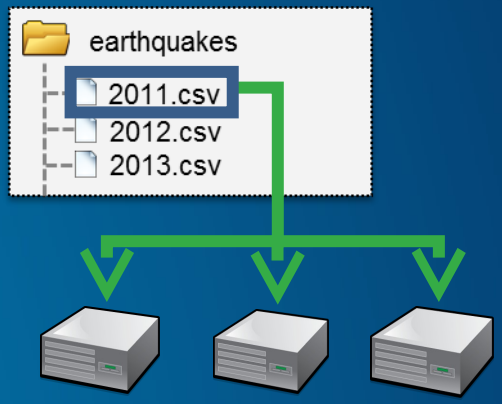
5 – 7 years old
Quad-core, 3GHz
16 GB RAM
1 TB fast disk

The Dredd Cluster

Apache Hadoop

Distributed Storage

- The Hadoop Distributed File System (HDFS) is a hierarchical file system where datasets are organized into directories and files
- These files are accessed like regular files, however they are actually distributed throughout the Hadoop cluster



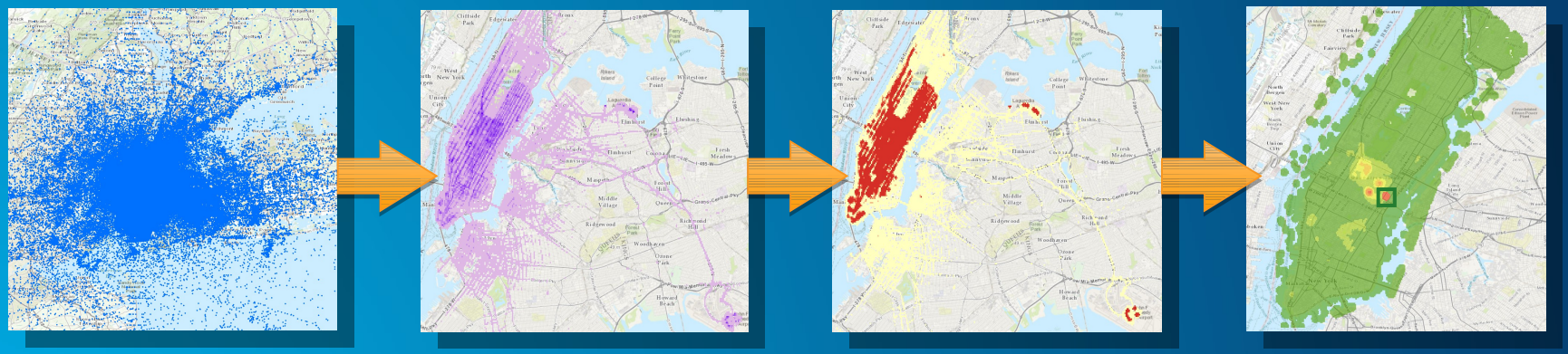
“Big data is not about the data.”
– Gary King
Harvard University
Director, Inst. For Quantitative Social Science

(Making the point that while data is plentiful and easy to collect, ***the real value is in the analytics****)*

GeoAnalytics for Big Data

Distributed analysis on distributed data

- Parallelized batch analytics on tabular, vector, raster, and imagery datasets (big and standard data)



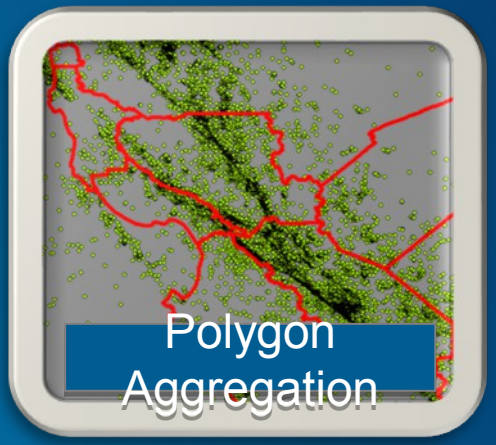
Raw Data | Aggregated Data | Hotspots | Analysis Results

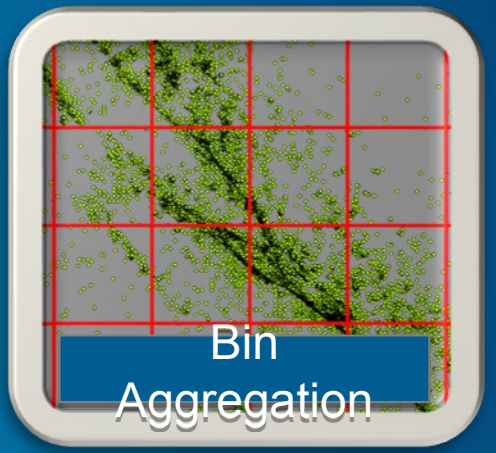
- Supports data exploration via feature, map, and image layers

Spatial Aggregation

Making Big Data Manageable

- Reduces the size of the data
- Maintains a summary of numeric attributes
- Common aggregation patterns
  - Polygons
    - Aggregate points into polygons
  - Bins
    - Aggregate points into bins defined by a grid





Spatial Aggregation

Who needs it?

I don’t need it. I’ll just draw everything.

FAA flight data – 17.5 million points

Spatial Aggregation

Why it's important

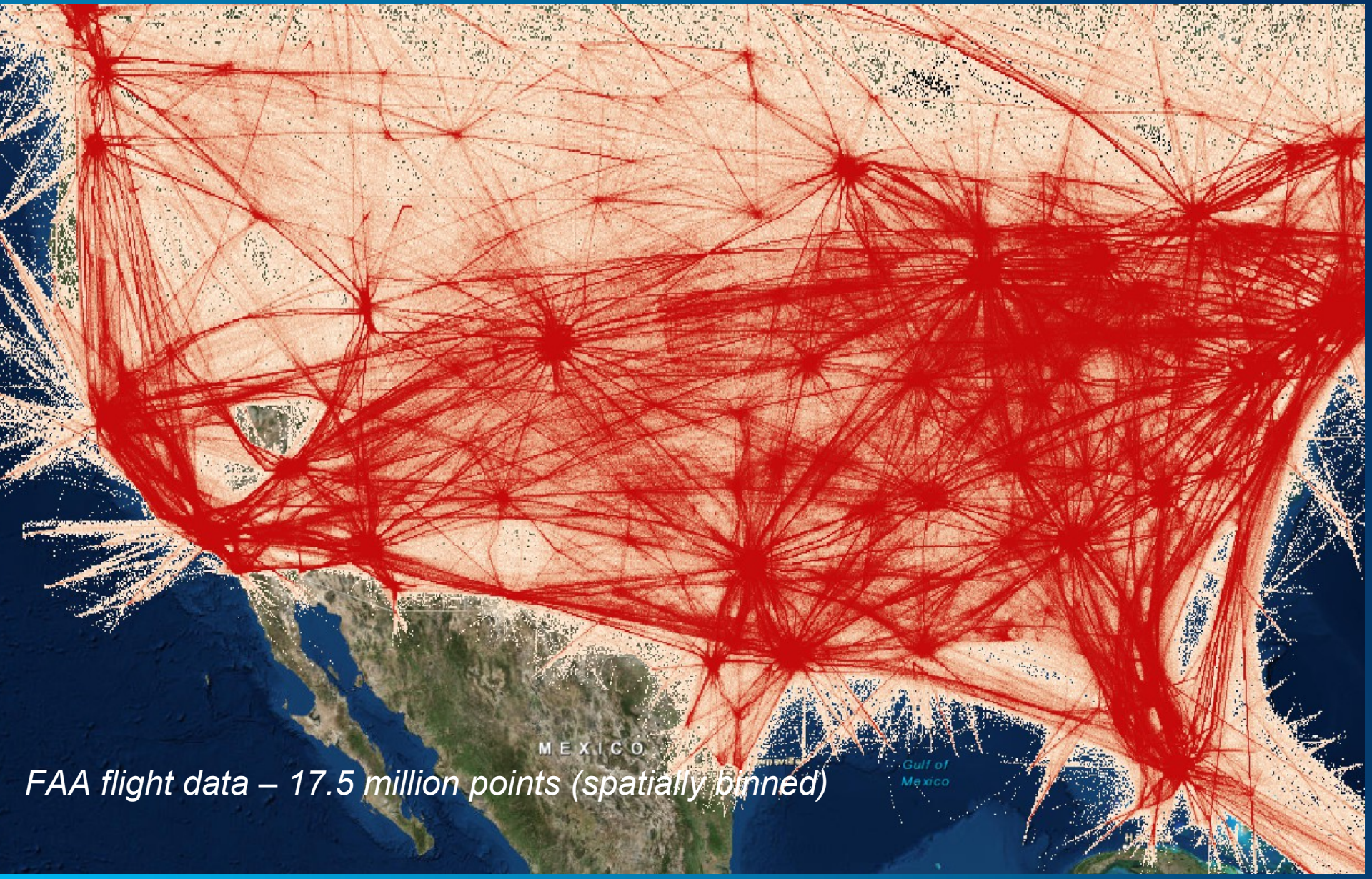
- **Mapping millions to billions of points in a small, dense area just isn't feasible**
  - **It's slow**
  - **It's difficult to identify patterns**

FAA flight data – 17.5 million points

Spatial Aggregation

Why it's important

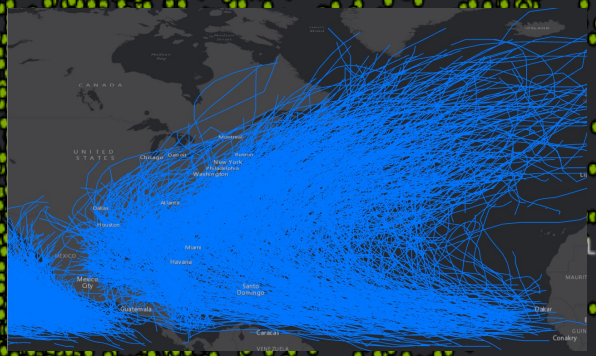
- Mapping millions to billions of points in a small, dense area just isn't feasible
  - It's slow
  - It's difficult to identify patterns

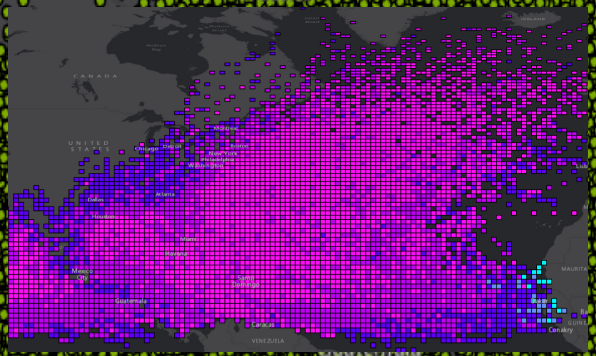


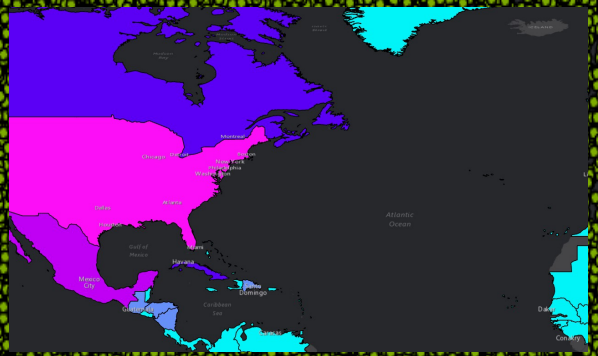
FAA flight data – 17.5 million points (spatially binned)

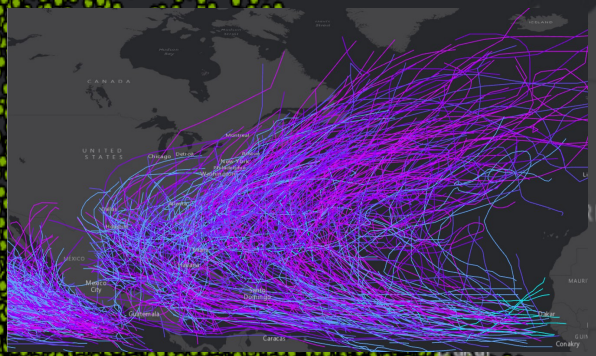
Aggregation

Summary methods









Precision Agriculture

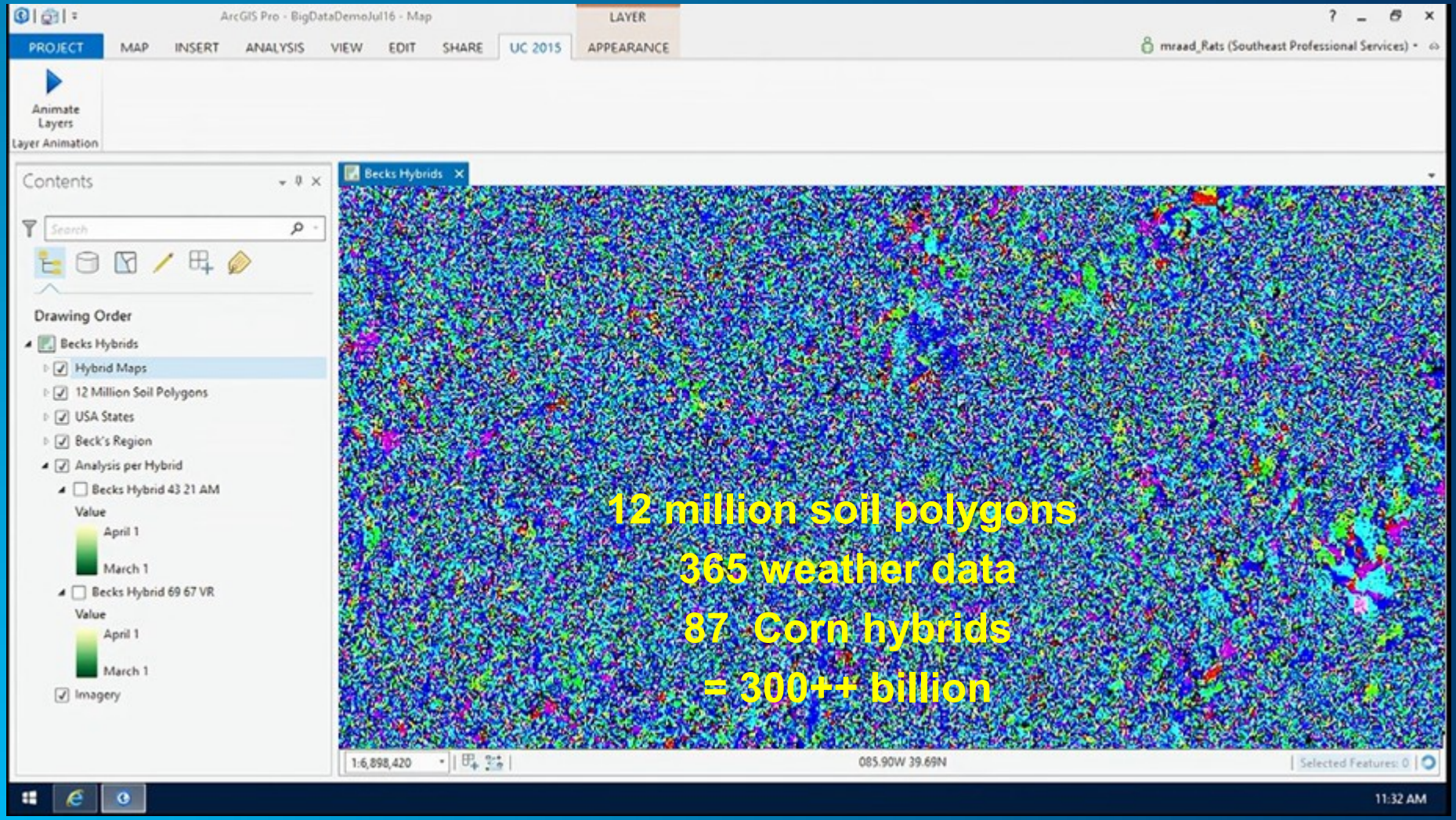
Beck's produces 87 varieties of corn hybrids

Where?

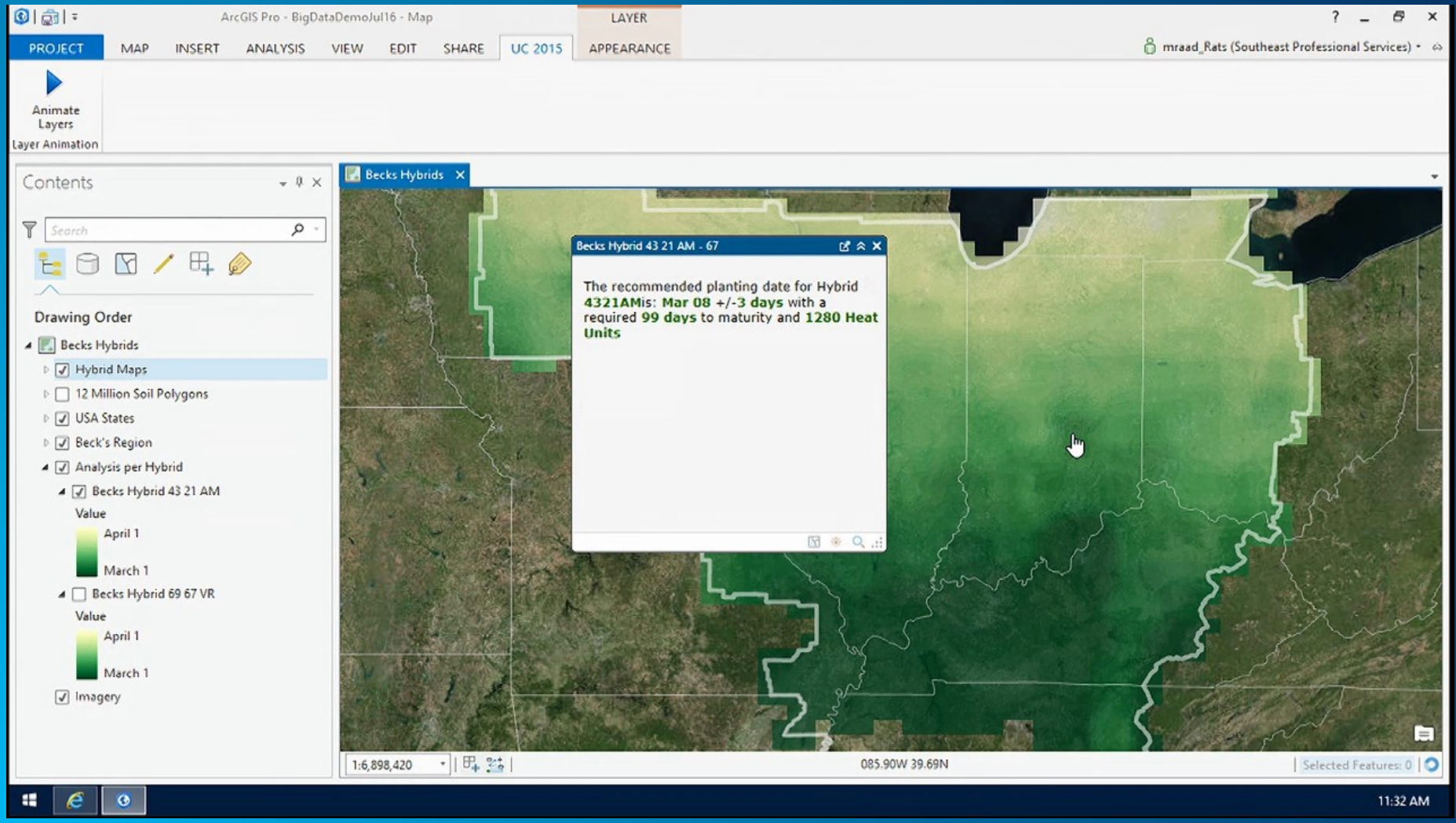
When?

Spatial and Temporal Calculations

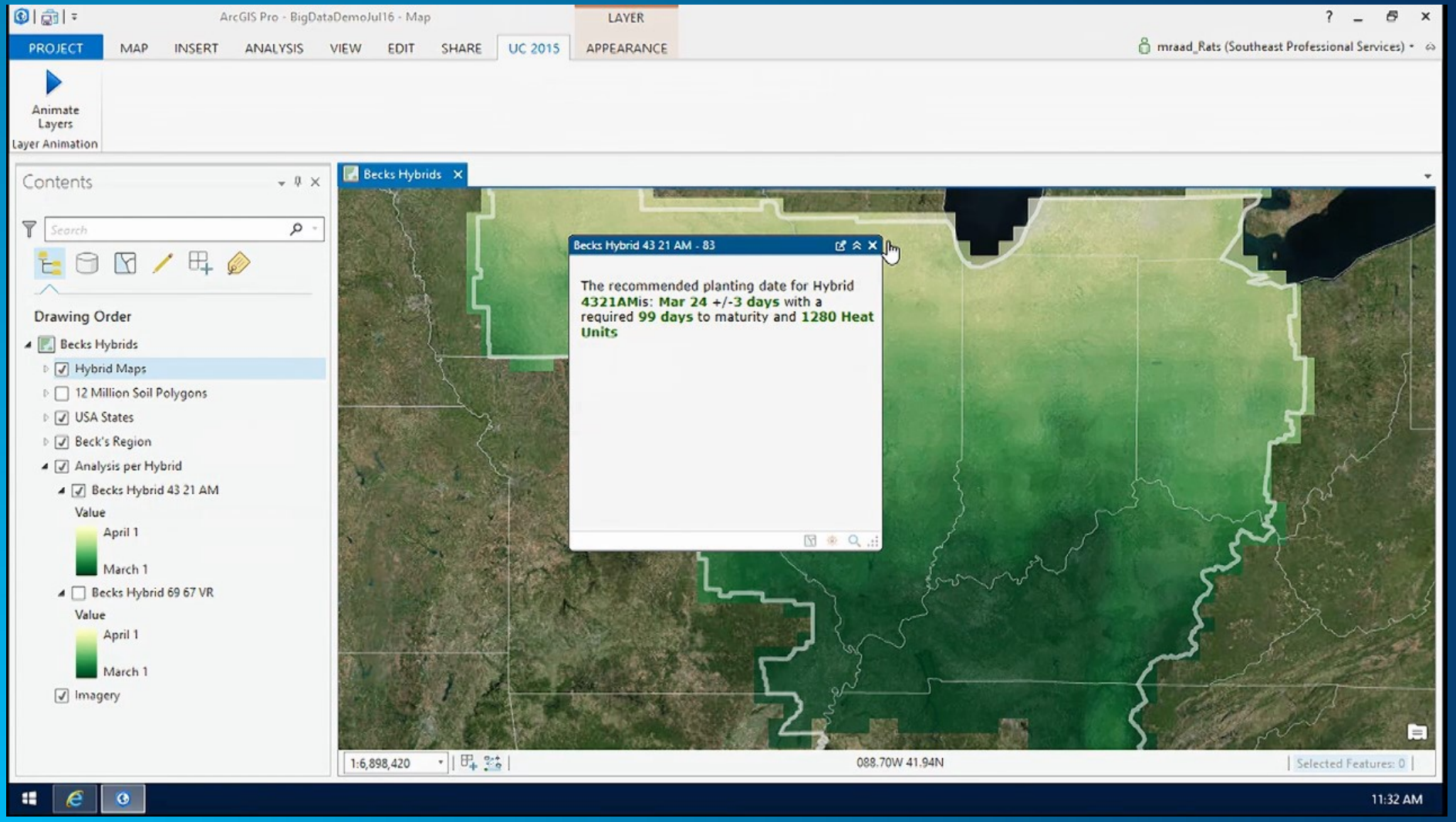
300 billions calculations in 10 minutes



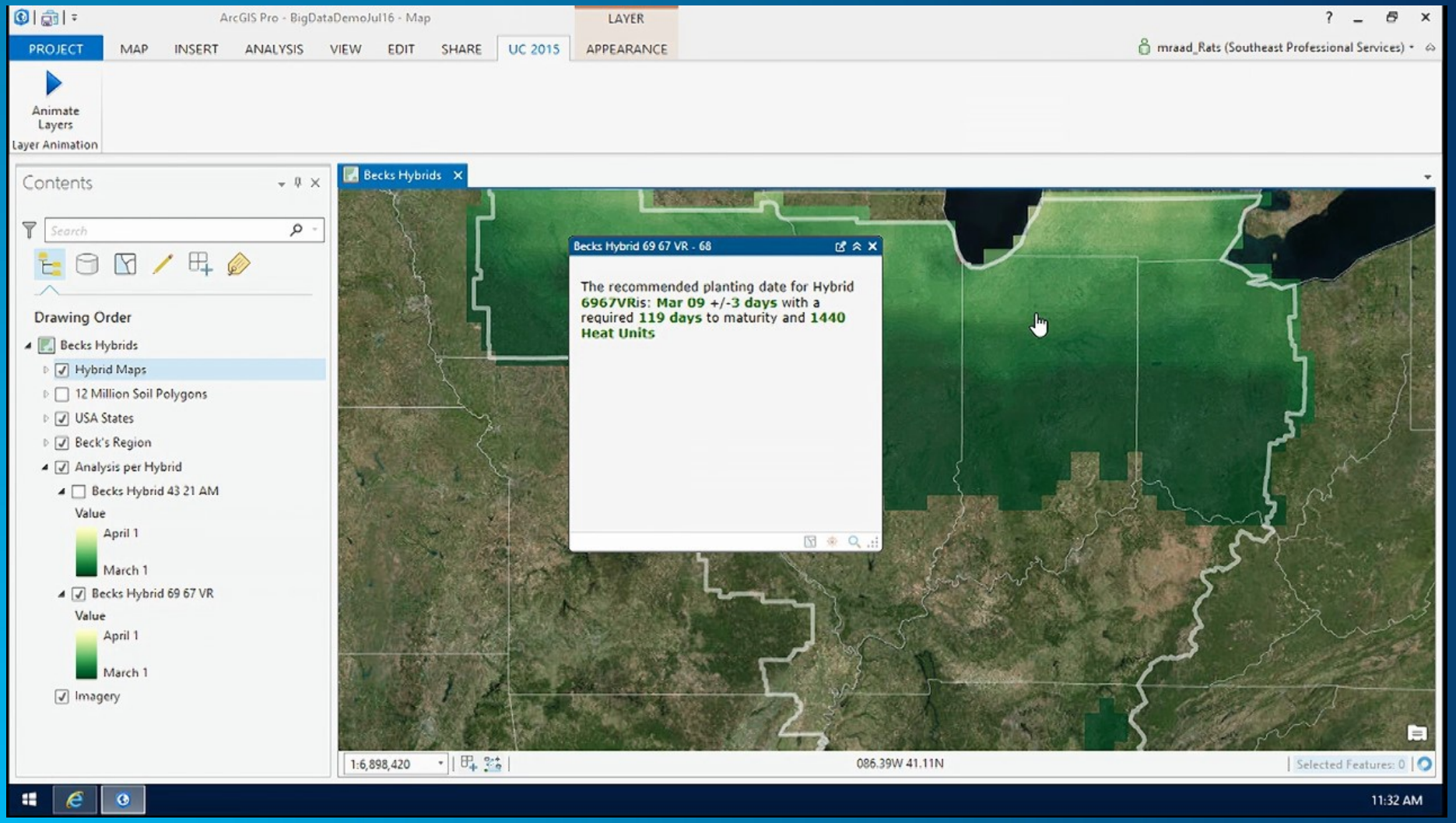
Hybrid 4321


ArcGIS Pro - BigDataDemoJul16 - Map
LAYER
PROJECT
MAP
INSERT
ANALYSIS
VIEW
EDIT
SHARE
UC 2015
APPEARANCE
mraad_Rats (Southeast Professional Services)
Animate Layers
Layer Animation
Contents
Search
Drawing Order
Becks Hybrids
Hybrid Maps
12 Million Soil Polygons
USA States
Beck's Region
Analysis per Hybrid
Becks Hybrid 43 21 AM
Value
April 1
March 1
Becks Hybrid 69 67 VR
Value
April 1
March 1
Imagery
Becks Hybrids
Becks Hybrid 43 21 AM - 67
The recommended planting date for Hybrid 4321AMis: Mar 08 +/-3 days with a required 99 days to maturity and 1280 Heat Units
1:6,898,420
085.90W 39.69N
Selected Features: 0
11:32 AM

Hybrid 4321


ArcGIS Pro - BigDataDemoJul16 - Map
LAYER
PROJECT
MAP
INSERT
ANALYSIS
VIEW
EDIT
SHARE
UC 2015
APPEARANCE
mraad_Rats (Southeast Professional Services)
Animate Layers
Layer Animation
Contents
Search
Drawing Order
Becks Hybrids
Hybrid Maps
12 Million Soil Polygons
USA States
Beck's Region
Analysis per Hybrid
Becks Hybrid 43 21 AM
Value
April 1
March 1
Becks Hybrid 69 67 VR
Value
April 1
March 1
Imagery
Becks Hybrids
Becks Hybrid 43 21 AM - 83
The recommended planting date for Hybrid 4321AMis: Mar 24 +/-3 days with a required 99 days to maturity and 1280 Heat Units
1:6,898,420
088.70W 41.94N
Selected Features: 0
11:32 AM

Hybrid 6967


ArcGIS Pro - BigDataDemoJul16 - Map
LAYER
PROJECT
MAP
INSERT
ANALYSIS
VIEW
EDIT
SHARE
UC 2015
APPEARANCE
mraad_Rats (Southeast Professional Services)
Animate Layers
Layer Animation
Contents
Search
Drawing Order
Becks Hybrids
Hybrid Maps
12 Million Soil Polygons
USA States
Beck's Region
Analysis per Hybrid
Becks Hybrid 43 21 AM
Value
April 1
March 1
Becks Hybrid 69 67 VR
Value
April 1
March 1
Imagery
Becks Hybrids
Becks Hybrid 69 67 VR - 68
The recommended planting date for Hybrid 6967VRis: Mar 09 +/-3 days with a required 119 days to maturity and 1440 Heat Units
1:6,898,420
086.39W 41.11N
Selected Features: 0
11:32 AM

Ship Tracking Application for Port of Rotterdam

Where are the ships?

Port of Rotterdam

AIS = Automatic Identification System

Radar and Control Stations

VTS – Vessel Tracking Services

Port of Rotterdam

Facts

8th Largest port in the world

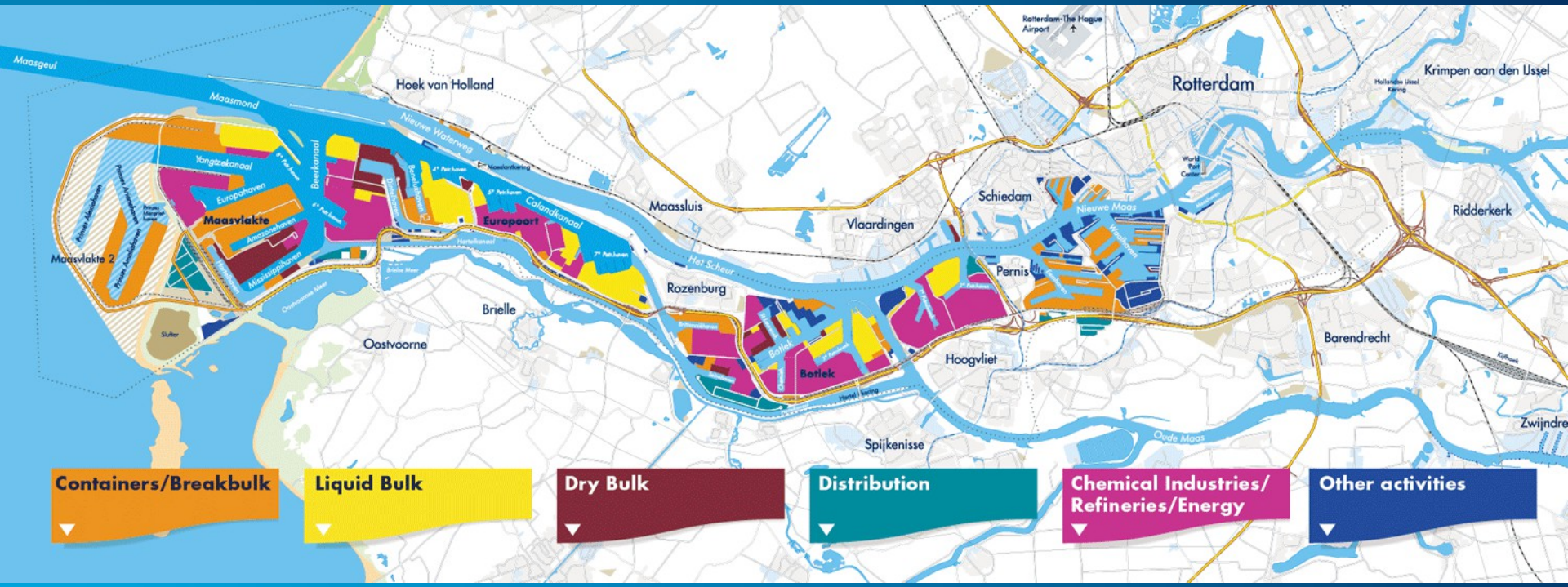
Largest port of Europe

Total area : 12,600 ha

Depth 24 meters

70.5 km quay length

Port of Rotterdam figures (1 years)



- 35.000 Ship visit with 400 million ton cargo
- 80.000 barge visit
- 7.500.000 trucks (25.000 per day)

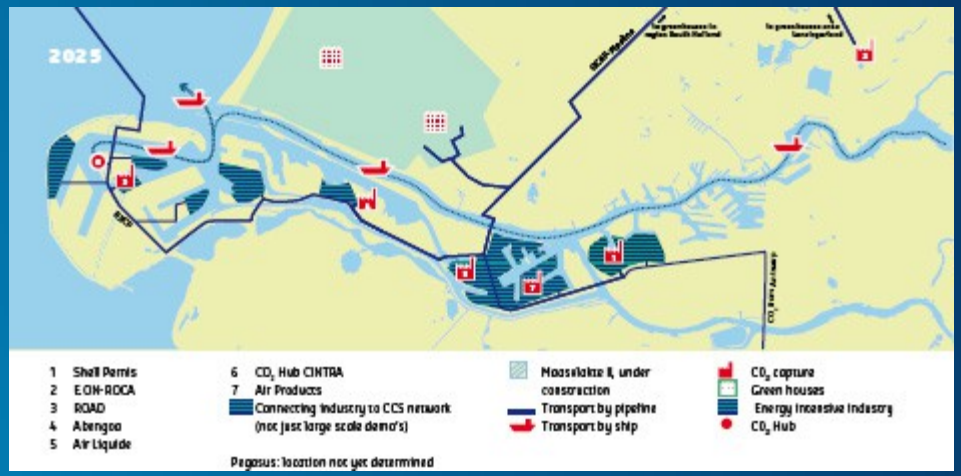
Usage of ship position data

Harbour master

- Incident Analysis
- Safety Checks

Capacity management

- Identifying Bottlenecks
- Planning Decision Supports



Environmental Management

- Pollution (NOx) Calculation
- Speed measures to reduce pollutions

Big Data

characteristics

- Volume
  - 18 billions records (since 2009)
- Velocity
  - >1000 records every 10 secs
- Format
  - CSV format

Consideration

Geospatial Database vs hadoop

Thank You