Presented at the FIG Congress 2018,
May 6-11, 2018 in Istanbul, Turkey

FIG 2018 ISTANBUL

06-11 MAY 2018
EMBRACING OUR SMART WORLD
WHERE THE CONTINENTS CONNECT:
ENHANCING THE GEOSPATIAL
MATURITY OF SOCIETIES

Management of Big Geographic Data for Smart City Applications

Prof. Dr. Arif Cagdas AYDINOGLU, Ress. Asst. Rabia BOVKIR

Geomatics Engineering Department, Gebze Technical University

Overview

- Introduction to the Smart City Concept
  - Requirements and application domains
  - Architecture
- Internet of things (IoT)
- Big Data concept
  - Components
  - Solutions
- Big Data Applications in Smart City Concept

Smart City Concept

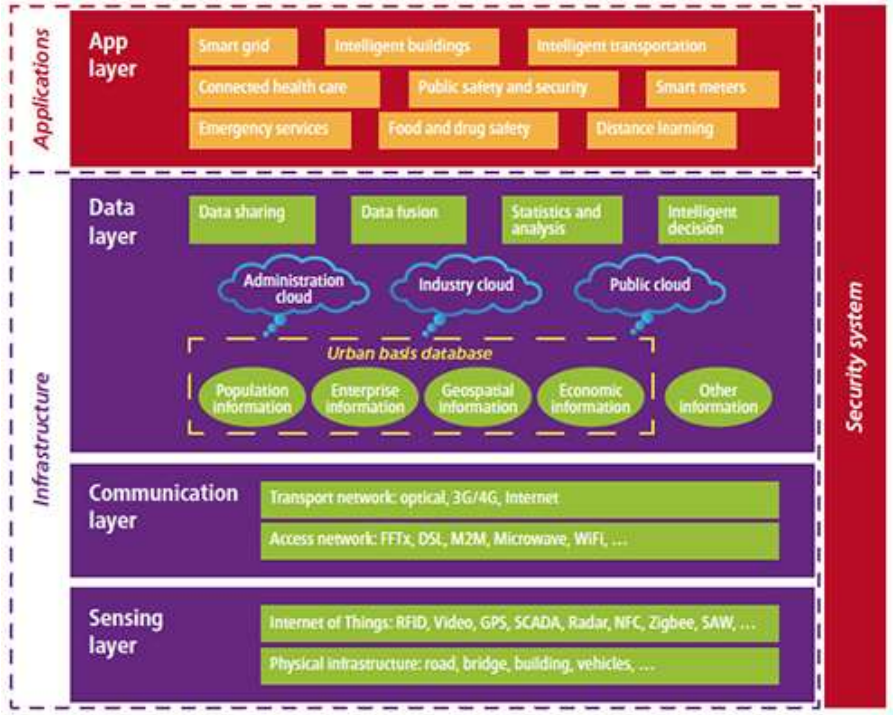
- Today, **more than 50%** of the world's population live in urban areas, and the reports published by the United Nations reveal that this number will reach 80% by 2050. As a result, **significant problems concerning traffic and transportation, energy, air and environmental pollution and emergency response** arise due to excessive population density.
- In this context, **"Smart City"** concept was introduced as a solution to all these problems. The concept can be expressed as the **modernization effort aiming to provide better service** to the city-dwellers by using all urban resources in an efficient and sustainable manner. In other words, it is the way of **monitoring and managing cities as a whole smarter system with the help of information technologies.**
- A smart city enables **efficient integration of physical, digital and human system** for providing a **sustainable, comfortable and inclusionary future** for its citizens.

Smart City Components

- In smart city concept, the main purpose is to provide **more quality and prospering life conditions** appreciated by its citizens through **Information and Communication Technologies (ICTs)** to enhance the utility of both urban and social services.
- Regardless of the application area, the smart city concept includes technology, data process and people. Considering the technology perspective, smart city involves four components:
  - Communication Interfaces includes **web services, portals, applications** for sending and receiving information from people and providers via open platforms.
  - Integrated Operation and Control Centers (IOCC) includes **technological** (computers, application monitors etc.), **physical** (operating and management rooms etc.) and **process infrastructures**, **government agency personnel** and **service providers** for processing and analysing the smart city issues.
  - Sensors and Connected Devices involves **capturing different kinds of signals** from the environment and **sending these signals to computers** in management centres via networks.
  - Connectivity Infrastructure involves **fixed and/or mobile data networks** which includes f**iber, cable and wireless** (Wi-Fi, 3G, 4G, or radio) transmission networks to send and receive data.

Smart City Architecture

- In a sustainable smart city architecture there are basically four layers:
  - Sensing layer involves **sensors and other data capturing devices** that are spread across the urban area which collect data about various thematic activities in the physical environment and send to data centres in the data layer.
  - Communication layer enables **the transmission and conversation between sensors and data processing** platform in the data layer.
  - Data layer can be accepted as the **intelligence layer of smart-city** architecture. This layer involves **data servers** that process data by applying different statistical models such as predictive, descriptive and decision.
  - Service/Application layer works as a cross-department operation centre. The **citizens can access and share the information via mobile applications and web portals** which are designed on this layer.

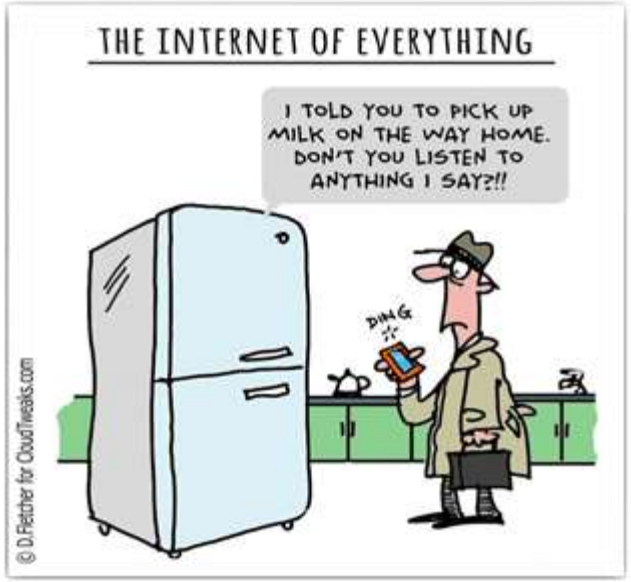


Source: Technical Report on ICT Infrastructure for Cyber-Security, Data Protection & Resilience.

Internet of Things (IoT)

- Today, data is generated at an ever-increasing rate due to the growing volume of social networking interactions, the increasing number **of location-sensitive devices** and **"smart sensors"** that capture and transmit information about the physical. With smartphones, even each individual can be considered as a sensor.

- All these developments bring out another important subject in the scope of smart cities: **"Internet of Things (IoT)"**. IoT can be defined as infrastructure based on interoperable information and communication technologies which makes advanced services possible by linking physical and virtual things/objects.



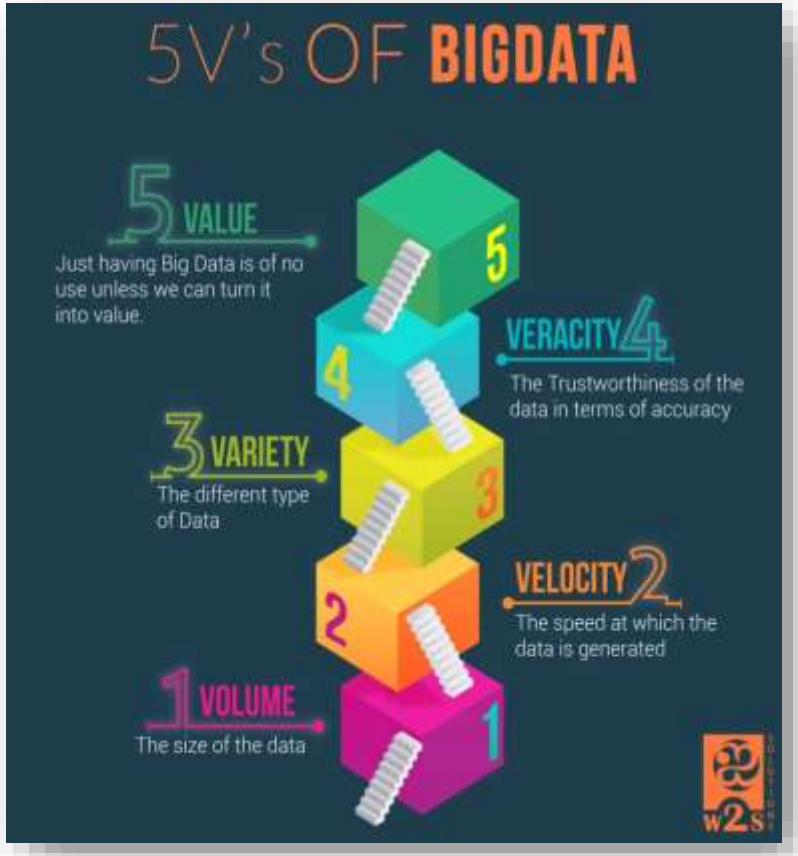
Big Data

- Along with the rapid growth and development of the technology, large volume of data become producable from smartphones, cameras, GPS, sensors, social media, games and applications.
- Due to the diversity and complexity of this large volume of data, processing this data effectively and rapidly is one of the most important questions in the smart city concept. Several definitions for big data concept are presented in the table .

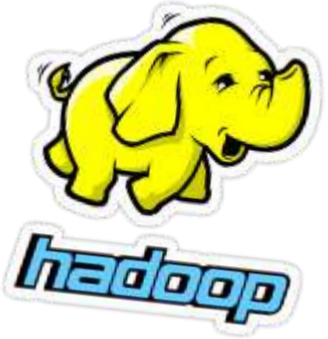
Source	Definition
TechAmerica, 2012	Big data is a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information.
Chen vd., 2012	A complete set of data sets and analytical techniques that are large and complex enough to require advanced and unique data storage, management, analysis and visualization technologies.
Schönberger and Cukier, 2013	The phenomenon that brings together three fundamental factors for analysing, understanding and organizing information: 1. More data, 2. Incomplete (unstructured) data, 3. Correlation.
Khan vd., 2014	Highly distributed and unstructured large volume data set that exceeds the processing capabilities of traditional database engines.
SAS, 2017	Big data is a popular term used to describe the exponential growth, availability, and use of information, both structured and unstructured.
IBM, 2017	Data coming from everywhere; sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction record, and cell phone GPS signal to name a few.
Oracle, 2017	Big data describes a holistic information management strategy that includes and integrates many new types of data and data management alongside traditional data.
Gartner Inc., 2017	Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.
NIST, 2017	Big Data consists of extensive datasets –primarily in the characteristics of volume, variety, velocity, and variability– that require a scalable architecture for efficient storage, manipulation, and analysis.
Microsoft, 2017	Big data typically refers to collections of datasets that, due to size and complexity, are difficult to store, query, and manage using existing data management tools or data processing applications.

Basic Principles of Big Data

- **Volume:** This concept refers to the **size of the data** that has been produced from all the sources. Because data producing techniques are increasing day by day, technologies for archiving, processing, integrating and storing such big data should be managed and planned properly.
- **Variety:** This component refers to **different types of data** being produced in different environments and generally not being structural.
- **Velocity:** This concept means **that production and distribution speed** of the big data is very high. Big data production speed increases every day and reaches an incredible dimension in seconds.
- **Value:** Value is considered as the most important component of the big data concept and means that **relevant data creates value** for applications, institutions, companies etc. In other words, large amount of data that are complex, diverse, and difficult to manage should be made available to the user.
- **Veracity:** Value of the big volume data depends on its **accuracy.** It is very important to make the necessary analysis for transition of wide range of large volume data through the right layers with right level of security and privacy.

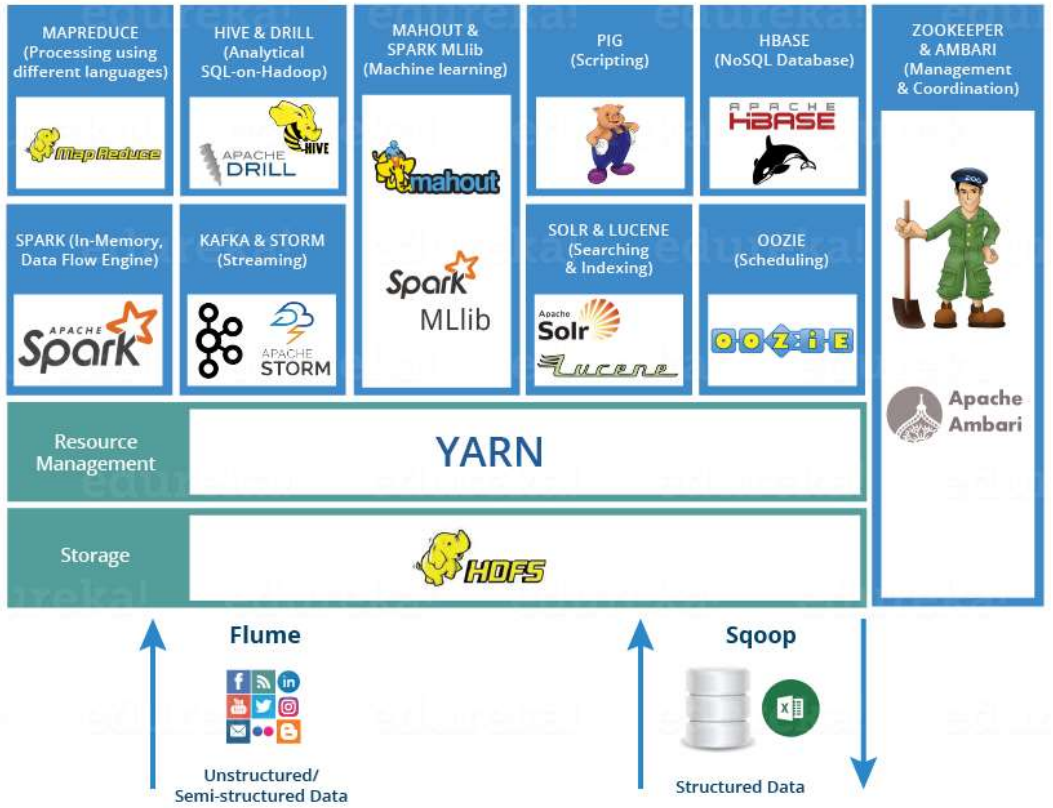


Big Data Solutions- Hadoop

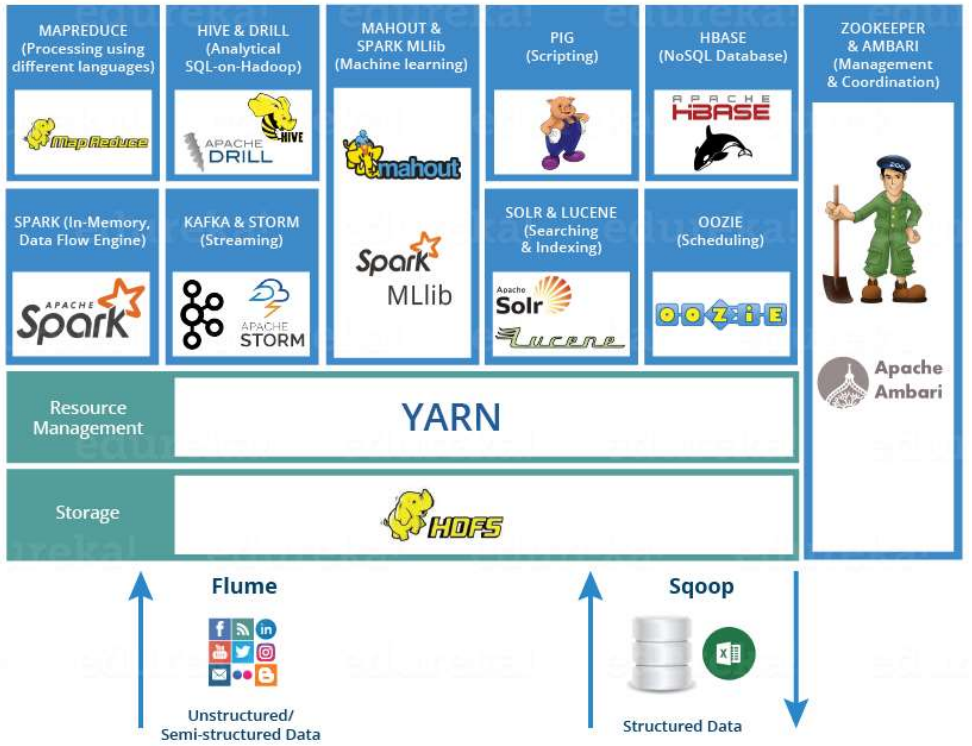


- Due to the characteristic features (volume, diversity, speed, etc.), **it is very difficult to benefit from big data** sets by using traditional processing methods. For this reason, various **methods, algorithms and technologies have been developing to process and benefit** from big data.
- One of the pioneering and most widely known big data technology is **Apache Hadoop.** The Apache Hadoop is a **open source software library** which **enables the distributed processing of large volume of data** across several computer clusters.
- Apache Hadoop includes the modules explained below:
  - Hadoop Common: The common utilities such as file system and operating system abstraction supporting the other Hadoop modules.
  - Hadoop Distributed File System (HDFS): A highly fault tolerant distributed file system which ensures high-throughput access to data.
  - Hadoop YARN: A framework for scheduling the jobs and managing the cluster resource.
  - Hadoop MapReduce: YARN-based system to process large volume of data that implements the MapReduce programming model in parallel clusters.

- The MapReduce supports the easy development of applications by enabling parallel process of large data sets stored in distributed architecture. The user uses map and reduce functions for defining a job and the system itself manages the required data partitioning, parallel processing and error management for this job. In this way large amounts of data can be read quickly and data processing speed and efficiency increase.
- The Hadoop ecosystem also provides an alternative execution engine to MapReduce called ***Tez.*** Tez is a generalized data-flow programming framework, built on Hadoop YARN, which provides a powerful and flexible engine to execute an arbitrary DAG of tasks to process data for both batch and interactive use-cases. Tez is being adopted by commercial software (e.g. ETL tools) and by other frameworks in the Hadoop ecosystem such as ***Hive*** and ***Pig***.



- ***Ambari***, a web-based tool for provisioning, managing, and monitoring Apache Hadoop clusters.
- ***Avro,*** is a data serialization system within the Apache Hadoop.
- ***Chukwa,*** is a data collection system for managing large distributed systems.
- ***Pig,*** is a high-level data-flow language and execution framework for parallel computation.
- ***ZooKeeper***, is a high-performance coordination service for distributed applications.
- ***Cassandra,*** is a scalable multi-master database with no single points of failure.
- ***HBase,*** is a scalable, distributed database that supports structured data storage for large tables.
- ***Hive,*** is a data warehouse infrastructure that provides data summarization and ad hoc querying.
- ***Drill,*** is a query engine that does not use MapReduce and support low latency queries natively on rapidly evolving multi-structure datasets at scale.
- ***Mahout,*** is a scalable machine learning and data mining library.
- ***Spark,*** is a fast and general compute engine for Hadoop data.



Big Data Solutions - NoSQL

- **NoSQL (not only Structured Query Language-SQL)** systems are emerged as an alternative to Relational Database Management System. A NoSQL database provides a mechanism for storage and retrieval of data that is modelled in means other than the tabular relations used in relational databases and allows searches from different structure and sizes of data.
- A common advantage of the NoSQL is that the **NoSQL databases scale horizontally** (or scale out), while the traditional databases scale up only. Most common NoSQL database types are;
  - **Key-Value Stores:** The Key-value (KV) stores are the simplest NoSQL databases that use the associative array as their fundamental data model. Examples of KV stores are the redis, Voldemort, riak, etc.
  - **Column Databases:** The column databases, data is stored as combinations of (key, value) pairs grouped into collections. Examples of popular column databases include the Hadoop databases HBase and Cassandra, Google databases BigTable and BigQuery, Impala etc.
  - **Document Stores:** In the document store concept, documents encapsulate and encode data (or information) in some standard formats such JSON and XML or encodings. Popular document stores include MongoDB, CouchBase, Apache CouchDB, eXist etc.
  - **Graph Databases:** A graph database is an online database management system with Create, Read, Update, and Delete (CRUD) methods that expose a graph data model. Examples of popular graph databases are the Neo4j, InfoGrid, Oracle Spatial and Graph, Oracle NoSQL Database etc.
  - **Array Databases**: The array databases provide database services specifically for arrays also called raster data. Arrays are used often to represent sensor, simulation, image, or statistics data. Most popular examples are, the Oracle Spatial and Graph, MonetDB, PostGIS, rasdaman, SciDB etc.

Big Data Solutions - NewSQL

- **NewSQL systems** seek to provide the same scalable performance of NoSQL systems for online transaction processing (OLTP) read-write workloads while still maintaining the ACID guarantees of a traditional database system.
- The NewSQL databases can be classified as:
  - **New architectures:** This type of systems are new database platforms and designed to operate in a distributed cluster of shared-nothing nodes, in which each node owns a subset of the data. Examples of NewSQL databases are *Google Spanner*, *Clustrix*, *VoltDB*, *MemSQL*, *SAP HANA*, *FoundationDB*, *NuoDB*, *ActorDB*, *Trafodion,* etc.
  - **SQL Engines:** The systems of this category are equipped with highly optimized storage engines for SQL, which provide the SQL programming interface. Examples include *TokuDB* and *InfiniDB.*
  - **Transparent sharding:** These systems provide a *database shard* which is a horizontal partition of data in a database or search engine, to split databases across multiple nodes automatically. Examples of the transparent sharding systems include *dbShards, Scalearc*, and *ScaleBase*.

Big Data Applications in the Smart Cities

Smart city component	Big Data Project	Location
Traffic and Environmental Issues	Stockholm has implemented smart management and applications to address the traffic and environmental issues. Half of a million entries of waste fractions, weights and locations were realized. Such a large amount of data was used to collect and analyze waste management collection data to identify the inefficiencies in waste collection routes within the city (Shahrokni et al., 2014). In Singapore, OneMap portal which can be openly accessed on the internet contains free up-to-date information POIs such as WiFi hotspots, public services, road conditions, national monuments etc. By using OneMap citizens can compute distances between potential targets or acces instant traffic information (Tao, 2013).	Sweden Singapore
Healthcare	Ministry of Health and Welfare in South Korea initiated the Social Welfare Integrated Management Network with the purpose of managing welfare benefits and services to citizens provided by the central and local government. Within the project 385 different types of public data from 35 agencies were analyzed (Kim et al., 2014).	South Korea
Safety	In order to address national security, infectious diseases, and other national concerns, the Singapore government launched the Risk Assessment and Horizon Scanning (RAHS) program in 2004. Within the Project large-scale data sets were analysed in order to manage national threats, including terrorist attacks, infectious diseases, and financial crisis (Kim et al., 2014).	Singapore

Big Data Applications in the Smart Cities

Smart city component	Big Data Project	Location
Natural Resources & Energy	The UK government established the Horizon Scanning Centre (HSC) in 2004 to improve the government's ability to deal with cross-departmental and multi-disciplinary challenges. In 2011, the HSC's Foresight International Dimensions of Climate Change effort addressed climate change and its effects on the availability of food and water, regional tensions, and international stability and security by performing in depth analysis on multiple data channels (Kim et al., 2014).	UK
Public Administration	With the Helsinki Region Infoshare more than 1030 databases which cover a wide range of urban phenomena such as transport, economics, conditions and employment were made available to individuals, government, academy, business and research institutions. The project promotes the opportunity of public involvement in decision-making process.	Finland
Government & Agency Administration	In 2009, the U.S. government launched data.gov as a step toward government transparency and accountability. It is a warehouse containing 420,894 datasets covering transportation, economy, health care, education, and human services and the data source (Neirrotti et al., 2014). In 2011, Syracuse, NY, in collaboration with IBM, launched a Smarter City project to use big data to help predict and prevent vacant residential properties. Michigan's Department of Information Technology constructed a data warehouse to provide a single source of information (Neirrotti et al., 2014).	USA

Conclusion

- Reviewing smart city applications show how big data analysis is important for successful urban management. Analysing and recognizing the patterns especially in traffic, healthcare and safety issues, providing detailed and immediate data solutions or integration and collaboration of government agencies and citizens in decision making helps to reach successful smart city.
- However, some characteristics especially in technological manner are still not well defined and assimilated in smart city concept. Right tools and methods, therefore, should be used for big data processing and for achieving successful applications and advance services in smart cities.
- In smart city applications, different data types and technologies are required according to the target application domain. Therefore, all aspects should be analysed and feasible investment strategies should be determined by decision makers.

Thanks for your attention…

Prof. Dr. Arif Cagdas AYDINOGLU
Gebze Technical University
Dept. of Geomatics Engineering 41400
Gebze/Kocaeli Turkey
Email: aydinoglu@gtu.edu.tr
Web: arifcagdas@com
Twitter:
https://www.twitter.com/arifcagdas

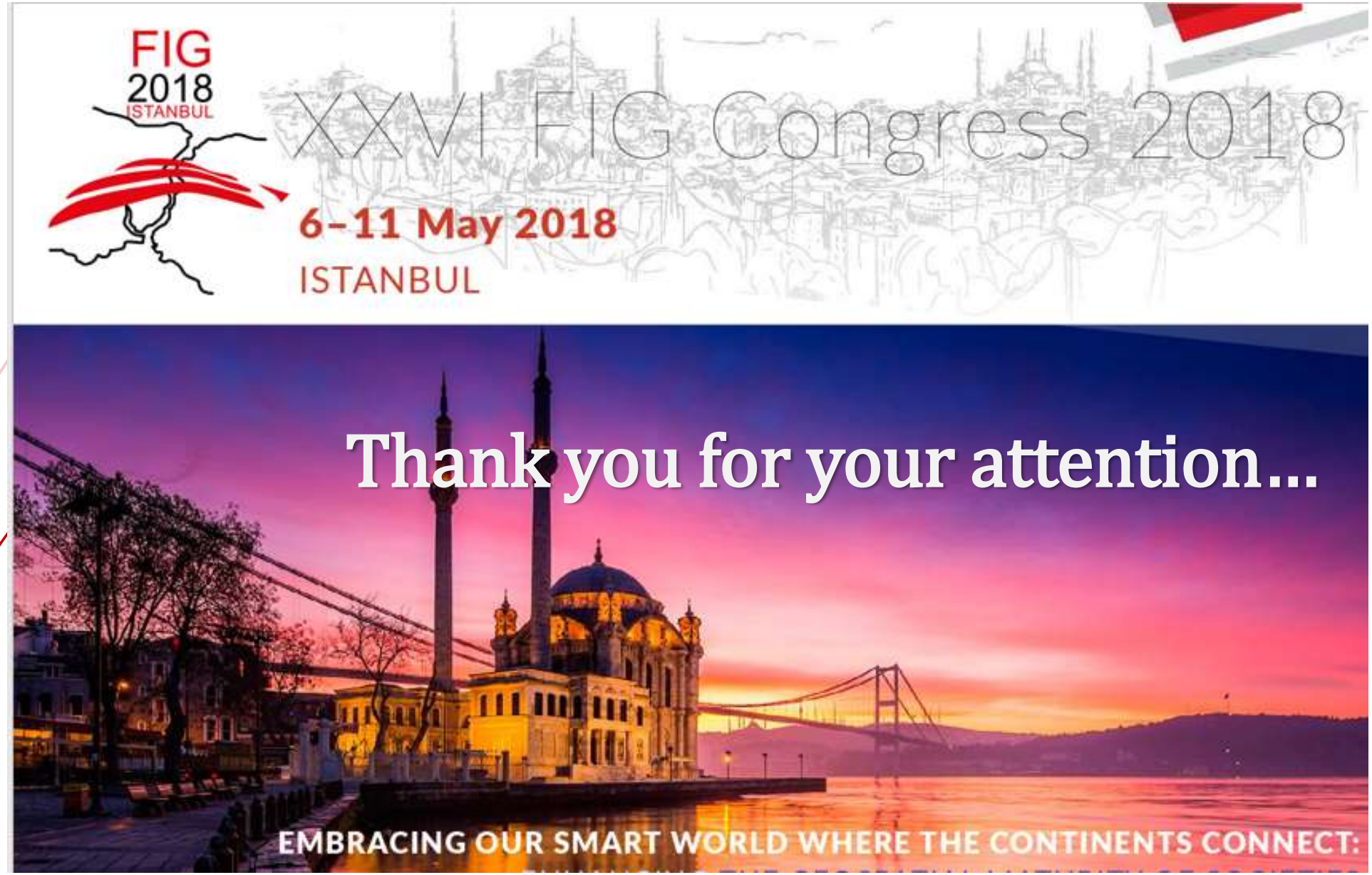


FIG
2018
ISTANBUL
XXVI FIG Congress 2018
6-11 May 2018
ISTANBUL
Thank you for your attention…
EMBRACING OUR SMART WORLD WHERE THE CONTINENTS CONNECT: